Natural Language Processing for Improving Textual Accessibility (NLP4ITA)

Workshop Programme

Sunday, May 27, 2012

9:00 – 9:10 Introduction by Workshop Chair

9:10 – 10:10 **Invited Talk** by Ruslan Mitkov
NLP and Language Disabilities

Session: Simplification

10:10 - 10:30 María Jesús Aranzabe, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios
First Approach to Automatic Text Simplification in Basque

10:30 - 11:00 Coffee break

11:00 - 11:20 Alejandro Mosquera, Elena Lloret, Paloma Moreda
Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalization Resources

Session: Resources

11:20 - 11:40 Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov
What can readability measures really tell us about text complexity?
12:00 - 12:20 Luz Rello, Ricardo Baeza-Yates, Horacio Saggion, Jennifer Pedler
A First Approach to the Creation of a Spanish Corpus of Dyslexic Texts

Session: Vocal Aid

12:20 - 12:40 Janneke van de Loo, Guy De Pauw, Jort F. Gemmeke, Peter Karsmakers, Bert Van
Den Broeck, Walter Daelemans, Hugo Van hamme
Towards Shallow Grammar Induction for an Adaptive Assistive Vocal Interface: a Concept Tagging Approach

12:40 - 13:00 Tiberiu Boroș, Dan Ștefănescu, Radu Ion
Bermuda, a data-driven tool for phonetic transcription of words

13:00 End of the Workshop — Lunch break

Editors

Luz Rello	Universitat Pompeu Fabra
Horacio Saggion	Universitat Pompeu Fabra

Workshop Organizers/Organizing Committee

Ricardo Baeza-Yates	Universitat Pompeu Fabra, Yahoo!
Paloma Moreda	Universidad de Alicante
Luz Rello	Universitat Pompeu Fabra
Horacio Saggion	Universitat Pompeu Fabra
Lucia Specia	University of Sheffield

Workshop Programme Committee

Sandra Aluisio	University of Sao Paulo
Ricardo Baeza-Yates	Universitat Pompeu Fabra, Yahoo!
Delphine Bernhard	University of Strassbourg
Nadjet Bouayad-Agha	Universitat Pompeu Fabra
Richard Evans	University of Wolverhampton
Caroline Gasperin	TouchType Ltd
Pablo Gervás	Universidad Complutense de Madrid
José Manuel Gómez	Universidad de Alicante
Simon Harper	University of Manchester
David Kauchak	Middlebury College
Guy Lapalme	University of Montreal
Elena Lloret	Universidad de Alicante
Paloma Martínez	Universidad Carlos III de Madrid
Aurelien Max	Paris 11
Kathleen F. McCoy	University of Delaware
Ornella Mich	Foundazione Bruno Kessler
Ruslan Mitkov	University of Wolverhampton
Paloma Moreda	Universidad de Alicante
Constantin Orasan	University of Wolverhampton
Luz Rello	Universitat Pompeu Fabra
Horacio Saggion	Universitat Pompeu Fabra
Advaith Siddharthan	University of Aberdeen
Lucia Specia	University of Sheffield
Juan Manuel Torres Moreno	University of Avignon
Markel Vigo	University of Manchester
Leo Wanner	Universitat Pompeu Fabra
Yeliz Yesilada	Middle East Technical University Northern Cyprus Campus

Table of contents

Author Index

Preface

In recent years there has been an increasing interest in accessibility and usability issues. This interest is mainly due to the greater importance of the Web and the need to provide equal access and equal opportunity to people with diverse disabilities. The role of assistive technologies based on language processing has gained importance as it can be observed from the growing number of efforts (United Nations declarations on universal access to information or WAI guidelines related to content) and research in conferences and workshops (W4A, ICCHP, ASSETS, SLPAT, etc.). However, language resources and tools to develop assistive technologies are still scarce.

This workshop Natural Language Processing for Improving Textual Accessibility (NLP4ITA) aimed to bring together researchers focused on tools and resources for making textual information more accessible to people with special needs including diverse ranges of hearing and sight disabilities, cognitive disabilities, elderly people, low-literacy readers and adults being alphabetized, among others.

NLP4ITA had and acceptance rate of 54%, we received 11 papers from which 6 papers were accepted. We believe the accepted paper are high quality and present mixture of interesting topics.

We would like to thank all people who in one way or another helped in making this workshop a success. Our special thanks go to Ruslan Mitkov for accepting to give the invited presentation, to the members of the program committee who did an excellent job in reviewing the submitted papers, to Sandra Szasz for designing and updating NLP4ITA website and to the LREC organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Luz Rello and Horacio Saggion
Barcelona, 2012

First Approach to Automatic Text Simplification in Basque

María Jesús Aranzabe*, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios****

IXA NLP Group, Basque Philology Department*, Languages and Information Systems** University of the Basque Country
Sarriena auzoa zg 48940 Leioa*, Manuel Lardizabal 1 48014 Donostia***
maxux.aranzabe@ehu.es, a.diazdeilarraza@ehu.es, igonzalez010@ikasle.ehu.es

Abstract

Analysis of long sentences are source of problems in advanced applications such as machine translation. With the aim of solving these problems in advanced applications, we have analysed long sentences of two corpora written in Standard Basque in order to make syntactic simplification. The result of this analysis leads us to design a proposal to produce shorter sentences out of long ones. In order to perform this task we present an architecture for a text simplification system based on previously developed general coverage tools (giving them a new utility) and on hand written rules specific for syntactic simplification. Being Basque an agglutinative language this rules are based on morphological features. In this work we focused on specific phenomena like appositions, finite relative clauses and finite temporal clauses. The simplification proposed does not exclude any target audience, and the simplification could be used for both humans and machines. This is the first proposal for Automatic Text simplification and opens a research line for the Basque language in NLP.

1. Introduction

Automatic Text Simplification (TS) is a NLP task which aims to simplify texts so that they are more accessible, on one hand, among others to people who learn foreign languages (Petersen and Ostendorf, 2007); (Burstein, 2009) or people with disabilities (Carroll et al., 1999); (Max, 2005). And, on the other hand, it is useful for advanced NLP applications such us machine translation, Q&A systems or dependency parsers (Chandrasekar et al., 1996). In either cases, it is of prime importance to keep the meaning of original text, or at least trying not to lose information.

TS systems and architectures have been proposed for languages like English (Siddharthan, 2006), Portuguese (Candido et al., 2009), Swedish (Rybing et al., 2010), and there is ongoing work for Arabic (Al-Subaihin and Al-Khalifa, 2011) and Spanish (Saggion et al., 2011). Considering the advantages that these systems offer, we will explain here the architecture for a TS system based on the linguistic approach done so far for the Basque language, an agglutinative free-order language, in which grammatical relations between components within a clause are represented by suffixes.

This paper is structured as follows: in section 2 we explain briefly the linguistic typology of Basque associated to our problem. After that, in section 3 we present the corpora we have used. In section 4 we explain the process to simplify we propose and after it our architecture in section 5. The syntactic simplification proposals of the phenomena we have treated will be explained in section 6 and in section 7 we will expose this process by means of an example. We will finish the paper with the conclusion in section 8.

2. Typology of Basque

Basque is not an Indo-European language and differs considerably in grammar from the languages spoken in surrounding regions. It is, indeed, an agglutinative head-final pro-drop isolated language. The case system is ergative-absolutive. Due to its rich morphology, we have to take into account the structure of words (morphological analysis) to achieve this simplification task.

Basque displays a rich inflectional morphology. Indeed, it provides information about the case (Absolutive, Ergative or Dative) on either synthetic or auxiliary verbs. Basque declarative sentences are composed of a verb and its arguments and they can contain postpositional phrases too. The inflected verb is either synthetic or periphrastic. The synthetic (*noa*) in (1) is only composed by a word and it contains all the lexical and inflective information. The periphrastic (*joan nintzen*) in (2) is composed, however, of two (or three) words: main verb with lexical and aspect information and auxiliary verb containing agreement morphemes, tense and modality (Laka, 1996).

(1) *Etxera* *noa*
House-ALL go-1SG.PUNTUAL
'I go home'

(2) *Etxera* *joan* *nintzen*
House-ALL go-PRF AUX-1SG
'I went home'

In order to build subordinating clauses we attach complementisers[1] (comp) to the part of the verb containing inflection information. After the complementiser *-(e)n* in (3)(it is both past and comp) suffixes can be attached *-(e)an*-INE)[2]

(3) *Etxera* *joan* *nintzenean*
House-ALL go-PRF aux-1SG.COMP.INE
'When I went home'

The canonical element order is Sub Dat Obj Verb, but it can be easily changed according to the focus. Adjuncts can be placed everywhere in the sentence and arguments are often elided (pro-drop). The order changes in negative sentences as well. Let us see the first sentence in negative in (4).

[1] In sense of a morpheme which introduces all types of subordinating clauses

[2] INE=inessive(locative), ALL=allative, PRF=perfective

(4) *Ez noa etxera*
not go-1SG.PUNTUAL House-ALL
'I'm not going home'

3. Corpora analysis

We have used two corpora for this task: EPEC: *Euskararen Prozesamendurako Errenferentzia Corpusa*-Reference Corpus for the Processing of Basque (Aduriz et al., 2006a) and *Consumer* corpus (Alcázar, 2005).
EPEC corpus contains 300 000 words written in Standard Basque and it is tagged at morphological, syntactical levels (dependency-trees) (Aranzabe, 2008), and semantic level: word senses according to Basque WordNet and Basque Semcor (Agirre et al., 2006) and thematic roles in (Aldezabal et al., 2010). It is being tagged too at the pragmatic level: discourse markers (Iruskieta et al., 2011) and anaphora (Aduriz et al., 2006b).
Consumer corpus[3] is used in machine translation since the texts it contains are written in four languages (Spanish, Basque, Catalonian and Galician). It is a specialised corpus, compliling texts published the *consumer* magazine: critics, product comparison and so on.
The main characteristic of those corpora is that they contain authentic text.
In order to study the structures that should be simplified in Basque, to get better results in advanced application such as machine translation, we have taken the longest sentences from both corpora. We based our hypothesis on the results obtained by the machine translation system developed in our group when translating sentences of different length (Labaka, 2010). The results show that, the longer sentence longer, the higher error rate in Basque Spanish translation (table 1). The error rate used for scoring the results is HTER (Human-targeted Translation Error Rate) (Snover et al., 2006).

Words per sentence	**0-5**	**0-10**	**10-20**	**>20**
Sentences in corpora	5	41	100	59
HTER	17,65	28,57	32,54	49,16

Table 1: Sentence length and error rate in MT

Taking into account the results of the analysis of both corpora, we show in table 2 the sentence number we have treated in the corpora analysis and number that should be simplified, since they are complex sentences (with one or more complementisers). The third and fourth lines show the number of words that the longest and the shortest sentences we have in both corpora.

4. Simplification Process

The simplification process illustrates the operations that should be done and the steps we follow in order to produce simple sentences out of long sentences. Some of the operations we make have already been proposed in other TS works for other languages (Siddharthan, 2006) and (Aluísio et al., 2008).
In what follows we explain the operations considered:

	EPEC	**Consumer**
Long sentences	595	196
Complex sentences	488	173
Words/longest sentence	138	63
Words/shortest sentence	14	22

Table 2: Number of sentences and sentence length in Corpora

1. **Spliting:** Make as many new sentences as clauses out of the original.

2. **Reconstruction:** Two operations take place:

   (a) Removing no longer needed morphological features like complementisers (comp). Being Basque an agglutinative language we have to remove parts of words and not a whole word in case of finite verbs.

   (b) Adding new elements like adverbs or paraphrases. The main goal is to maintain the meaning.

3. **Reordering:** Reorder the elements in the new sentences, and ordering the sentences in the text.

4. **Adequation and Correction:** Correct the possible grammar and spelling mistakes, and fix punctuation and capitalisation.

This process will be ilustrated in section 7 by means of an example.

5. System Architecture

In this section we will present the architecture of the system we propose (see figure 1) to perform the steps mentioned in section 4. Having as input the text to be simplified, we distinguish different steps in our process:

1. The first step will be to evaluate the complexity of the text by means of a system already developed by our group for the auto-evaluation of essays *Idazlanen Autoebaloaziorako Sistema (IAS)* module (Castro-Castro et al., 2008). This module examines the text in order to determine its complexity based on several criteria such as the clause number in a sentence, types of sentences, word types and lemma number among others.

2. Once a sentence has been categorised as complex in the previous step, *Mugak* module (a system created in our group for detecting chunks and clauses) (Arrieta, 2010) will help us in the task of splitting long sentences into simple ones. *Mugak* is a general purpose clause identifier that combines rule-based and statistical-based clause identifiers previously developed for Basque. It works on the basis of the output produced by several tools implemented in our group[4]:

[3] http://corpus.consumer.es/corpus/

[4] http://ixa.si.ehu.es/Ixa

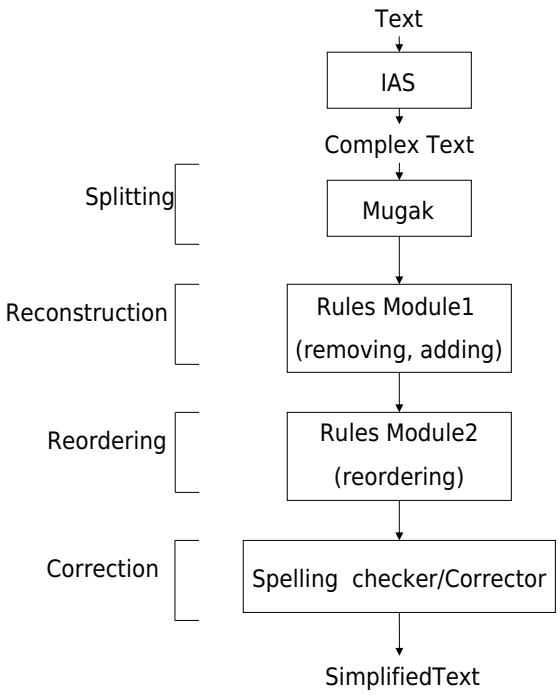


Figure 1: The architecture of system

- **Morpho-syntactic analysis:** *Morpheus* (Aduriz et al., 1998) makes word segmentation and PoS tagging. Syntactic function identification is made by *Constraint Grammar* formalism (Karlsson et al., 1995).
- **Lemmatisation and syntactic function identification:** *Eustagger* (Aduriz et al., 2003) resolves the ambiguity caused at the previous phase.
- **Multi-words items identification:** The aim is to determine which items of two or more words are always next to each other (Ezeiza, 2002).
- **Named entity recognition:** *Eihera* (Alegria et al., 2003) identifies and classifies named-entities in the text (person, organisation, location).

3. DAR (Deletion and Addition Rules) module includes a set of rules to perform the necessary deletions of morphological features and additions of grammatical elements in the split sentences. For example figure 2, shows the rule that would be applied to an auxiliary verb (aux) with a suffix in inessive, we remove the complementiser and the suffix (ine) and we add the adverb *ordu*-INE:

   We are defining the basic rules for the treatment of the phenomena explained in this paper. We are testing 15 rules and this process will be enriched while we go forward in our linguistic research.

4. ReordR (Reordering Rules) module includes a set of rules to perform the reordering needed in the created new sentences.

```
 if aux comp +ine {
remove comp and ine;
add ordu+ine in main clause;
}
```

Figure 2: A rule for an adverbial temporal sentences

5. Finally, the spell checker for Basque Xuxen (Agirre et al., 1992) will be applied in order to correct the created sentences.

6. Treated Phenomena

In the following subsections we give examples of the structures we have analysed and after them we give their simplifications. We follow the order that this structures have been explained in (Specia et al., 2008), i.e. apposition, relative clauses, adverbial subordinated clauses, coordinated clauses, non-inflected verb clauses and passive voice. In this paper we explain the simplification procedure for three structures: i) apposition and parenthetical structures, ii) finite relative clauses and iii) finite adverbial temporal clauses.
These structures are analysed in more details in (Gonzalez-Dios and Aranzabe, 2011).

6.1. Apposition and parenthetical structures

These structures give additional information about something that has been previously mentioned. Following we explain in (5) and (6) the process proposed for these structures. Sentences correspond to real text but have been shortened for clarity.
The steps for the treatment of (5) are:

1. When splitting we take the nominal group (NG) and the apposition to make several clauses out of the original one. In (5) NG are *Jose Maria Aznar* and *Javier Arenas* and their corresponding appositions are *Espainiako presidenteak* and *PPko idazkari nagusia*.

2. (a) We remove the apposition out of the original sentence.

   (b) Then, we add the copula verb to nominal group and the apposition, and so a new sentence is built (as we have here two apposition, two sentences will be built).

3. To reorder the elements in the sentence that has been built we follow this pattern:

```
NG(subj) apposition(pred) copula
```

   The ordering of the new sentences will be according to the order the appositions had in the original sentence (b) and (c) but the main clause in the original sentence will be the first one (a).

4. To check that the new sentences are grammatically correct and fix the punctuation by means of *XUXEN*.

(5) *Pankarta eraman zuten, besteak beste, Jose Maria Aznar Espainiako presidenteak eta Javier Arenas PPko idazkari nagusiak.*

The President of Spain Jose Maria Aznar and the Secretary-general of PP Javier Arenas carried the placard among others.

And those are the simplified sentences (a), (b) and (c):

a. *Pankarta eraman zuten, besteak beste, Jose Maria Aznarrek, eta Javier Arenasek.*

Jose Maria Aznar and Javier Arenas, carried the placard among others.

b. *Jose Maria Aznar Espainiako presidentea da.*

Jose Maria Aznar is President of Spain.

c. *Javier Arenas PPko idazkari nagusia da.*

Javier Arenas is Secretary-general of PP.

For parenthetical structures (6), we should repeat the process explained before. Sometimes we should retrieve the previously mentioned information as well to replace an elided element.

(6) *Hala ere, badirudi Sabino (Badajozetik fitxatuta), Moha (Barcelona B-tik) eta Aitor Ocio (Athleticek utzita) ez direla aurtengo fitxaketa bakarrak izango.*

However, it seems that Sabino (signed up from Badajoz), Moha (from Barcelona B) and Aitor Ocio (transferred from Athletic Bilbao) are not going to be the only signings.

And those are the simplified sentences (a), (b), (c) and (d):

a. *Hala ere, badirudi Sabino, Moha, eta Aitor Ocio ez direla aurteko fitxaketa bakarrak izango.*

However, it seems that Sabino, Moha and Aitor Ocio are not going to be the only signings.

b. *Sabino Badajozetik fitxatua da.*

Sabino is signed up from Badajoz.

c. *Moha Barcelona Btik fitxatua da.*

Moha is signed up from Barcelona B.

d. *Aitor Ocio Athleticek utzita da.*

Aitor Ocio is transfered from Athletic.

By simplifying the appositions this way the meaning of several entities will be *ipso facto* explained. Anyway, it would be necessary to explain the other entities in sentences, which are not appositions, if our target audience were humans (foreigners, second language learners, people lacking general knowledge). Sentences similar to the one presented here (with named-entities, references to persons, places etc.) could be enriched by facilitating access to Wikipedia[5]. This could be useful in a future proposal.

6.2. Relative clauses

Contrary to other subordinated clauses, relative clauses modify a noun and not a verb. There are different relativisation strategies in Basque: ordinary embedded relative clauses and appositive and extraposed relatives with relative pronouns (Oiarzabal, 2003). We consider that both can be simplified the same way. Sentence (7) is an example of the first strategy (ordinary embedded).

1. We split the sentence into relative clause and main clause. *Mugak* produces this output.
   (a) We will remove the complementiser.
   (b) We will copy the substantive they modified (the antecedent). In (7) the antecedent is *Ollanta Moises Humala teniente koronelak*. We will add the substantive to the previously removed relative clause, in the place of PRO[6], building a new simple sentence. We have to take into account the inflection case that the antecedent will have in the new sentence and give it the case that PRO has. If the clause is introduced by a relative pronoun, we use its inflection.
2. The subordinated clause will be left as it was, after having removed the complementiser.
3. To order the sentences we will keep the order they have in the original (relt (a) + main (b)).

This sentence (7) also presents an apposition linked to *Alberto Fujimori*, so in this case the treatment defined for appositions should be applied (here we just focused on finite relative clauses).

(7) *JOAN den igandeaz geroztik Alberto Fujimori Peruko presidentearen aurka altxamendu militar bat gidatzen ari den Ollanta Moises Humala teniente koronelak ez du uste bakarrik dagoenik (...)*

Since last Sunday Lt. Cr. Ollanta Moises Humala who is leading a military uprise against Peru president Alberto Fujimori does not think that he is alone.

And those are the simplified sentences (a) and (b):

a. *Joan den igandeaz geroztik Alberto Fujimori Peruko presidentearen aurka altxamendu militar bat gidatzen ari da Ollanta Moises Humala teniente koronela.*

Since last Sunday Lt. Cr. Ollanta Moises Humala is leading a military uprise against Peru president Alberto Fujimori.

[5] http://eu.wikipedia.org/wiki/Azala

[6] Phonetically null but syntactically active element

b. *Ollanta Moises Humala teniente koronelak ez du uste bakarrik dagoenik (...)*

Lt. Cr. Ollanta Moises Humala does not think that he is alone.

This will be the simplification of the most common finite relative clause type in Basque.

6.3. Adverbial temporal clauses

Adverbial clauses are adjuncts that specify relations like time, place, cause, consequence...with a reference to a main verb. As they constitute a heterogeneous group, we have decided to begin our experiment with the finite temporal adverbial clauses, and in the future we will expand our research.

1. So, we will split the original sentence (8).

2. The original main sentence will only be changed by adding an adverb (in (8) *orduan*) and by removing the subordinated clause. The subordinated will be left as the original, after having removed the complementiser and the suffix, which are attached to the auxiliary verb in case of periphrastic verbs, or to the main verb if the verb is synthetic.

   The element we add will be built this way: *ordu*-SUFFIX. The suffix is the one that is in the verb of the subordinated clause after the complementiser.

3. The problem with these clauses will be the ordering of new sentences and it will be more problematic if there are anaphoric elements. Meanwhile we have decided to keep the order the clauses in the original sentence, and if there is more than a subordinate clause, to put the former subordinated before the main clause, when they become simple sentences. In (8) both ordering have the same effect (a) and (b).

4. The new sentences will be corrected, if necessary, and punctuated.

(8) *erabakia hartu behar izan zuenean, ez zuen inolako zalantzarik izan don Polikarpo Gogorzak.*

'When he/she/it needed to decide, Sir Polikarpo Gogorza had no doubt.'

The simplified sentences are (a) and (b):

a. *Erabakia hartu behar izan zuen.*

   'He/she/it needed to decide.'

b. *Orduan, ez zuen inolako zalantzarik izan don Polikarpo Gogorzak.*

   'Then/in that time Sir Polikarpo Gogorza had no doubt.

We think that the procedure we have presented here will be useful for other adverbial clauses.

7. Example

We will explain here the process explained in section 4. Sentence (9) has the three phenomena we have presented in this paper. The changes we want to point out are underlined. We use the glosses in order to illustrate the morphological process properly, when needed.

Let us explain some morpho-syntactic aspects of the sentence (9) before showing the simplification steps:

There are 5 verbs in sentence (9), and each one builds a clause. The main verb is *da*, therefore it builds the main clause. The verb *dute* is main too, but in our analysis system it is dependent on the substantive it is referring to as apposition. The periphrastic verbs *igurtzitzen ditugunean* and *sortzen den* build subordinated clauses, and contrary to *gertatu* they are inflected. The non-inflected verb *gertatu* will be simplified although it is not treated in this approach. It will be treated when we treat non-inflected verbs.[7]

1. **Spliting:** Each verb forms a clause and they will be separated from the original one.
   Temporal adverbial clause: *(S Metalak igurtzitzen ditugunean S)*
   Non-finite verb concessive clause: *(S nahiz_eta kargen bereizketa berdin gertatu S)*
   Relative clause: *(S sortzen den S)*
   Main clause: *(S partikulen mugimendua oso erraza da material hauetan S)*
   Apposition: *(S eroankortasun elektriko haundia dute S)*

2. **Reconstruction:** Two steps are performed:

   (a) **Removing:** The complementisers (-(e)n) and suffixes in subordinated clauses (-(e)an).
   (S Partikulen mugimendua sortzen da s)
   (S Metalak igurtzitzen ditugu S)
   (S sortzen da S)

   (b) **Adding:** Adverbs and nominal groups in simple sentences.
   (S Orduan partikulen mugimendua oso erraza da material hauetan S)
   material hauek (S eroankortasun elektriko haundia dute S)

3. **Reordering:** This step is not needed in this sentence.
   (S Metalak igurtzitzen ditugu S)
   (S partikulen mugimendua sortzen da S)
   (S Orduan nahiz_eta kargen bereizketa berdin gertatu, partikulen mugimendua oso erraza da material hauetan S)
   (S material hauek eroankortasun elektriko haundia dute S)

[7] IMPF=imperfective, GEN=genitive, ERG=ergative ABS=Absolutive

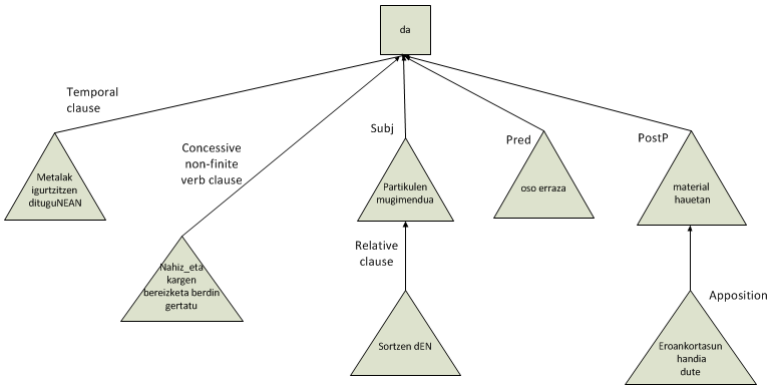


Figure 3: Tree of original sentence in example (9)

(9) *Metalak* *igurtzitzen* *ditugunean,* *nahiz_eta* *kargen* *bereizketa* *berdin*
Metal-ABS.PL rub-IMPF aux-ABS3PL.ERG1PL.COMP.INE although charge-GEN separation-ABS equal
gertatu, *sortzen* *den* *partikulen* *mugimendua* *oso* *erraza* *da*
happen-PRF create-IMPF aux-ABS3SG.COMP(REL) particle-GEN movement-ABS grad easy-ABS is
material *hauetan* *(eroankortasun* *elektriko* *haundia* *dute).*
material det-INE conductivity-ABS electrical big have.

'When we rub metals, although charge separation happens equally, the particle movement that is generated is very easy in these materials (they have a high electrical conductivity).

4. **Correction and Adequation:**

Correct sentences can be seen glossed in (10) and the trees in figure 4. Sentences have been punctuated and a non standard verb *igurtzitzen* and a non standard adjective *haundia* have been corrected (standardised) in this step.

(10) a. *Metalak* *igurzten* *ditugu.*
Metal-ABS.PL rub-IMPF auxABS3PL.ERG1PL
'We rub metals.'

b. *Partikulen* *mugimendua* *sortzen*
Particle-GEN movement-ABS generate-IMPF
da.
aux-3SG
'The particle movement is generated'

c. *Orduan,* *nahiz_eta* *kargen*
Then(hour-INE) although charge-GEN
bereizketa *berdin* *gertatu,*
separation-ABS equal happen-PRF
partikulen *mugimendua* *oso*
particle-GEN movement-ABS grad
erraza *da* *material* *hauetan.*
easy-ABS is material det-INE
'Then although charge separation happens equally, the particle movement is very easy in these materials.'

d. *Material* *hauek* *eroankortasun*
conductivity-ABS electrical big
handia *dute.*
have
'These materials have a high electrical conductivity.'

At the end of the simplification process, the tree in figure 3 becomes 4 trees that we can see in figure 4. The inserted elements are ovals, main verbs are squares, and other constituents are triangles.

8. Conclusions

In this paper we have presented an approach for building a TS system for the Basque language, proposing an architecture and explaining simplification proposals for apposition and parenthetical structures, finite relative clauses and finite temporal clauses.
The approach is based on the linguistic study we have performed on long sentences taken from two corpora (EPEC and Consumer).
Similarly to other studies (Specia et al., 2008) our analysis leads us to detect the sentence structures susceptible of being simplified.
Although our first motivation was to produce simple sentences to help in advanced applications such as machine translation, we think that this study is valid for other purposes: education, foreign language learners and so on.
Most of the tools that are proposed in this work have been developed for general purposes and we are reusing them. Besides, we have evaluated them while we looked at the

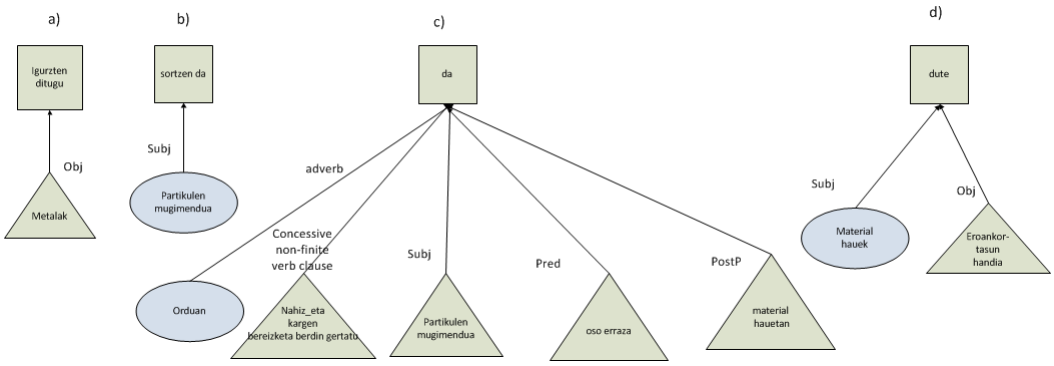


Figure 4: Tree of simplified sentences in example (10)

way to adapt them for our purpose. In this evaluation process we have concluded that *IAS* and *Mugak* are useful and that they can be a module of our architecture.

In any case, applying these rules we propose we get shorter sentences (Gonzalez-Dios and Aranzabe, 2011), which are translated automatically more easily, without losing the original meaning.

Although we have focused on syntactic simplification in this approach, it is important not to forget that in the future we should work on lexical simplification and text adaptation like proposed in (Siddharthan, 2006). We should remark as well that a part of this syntactic simplification approach is based on morphological constituents, which is necessary for high inflection languages like such a Basque. It is important to mention too that the operations and the steps we make are similar to those which are made in other languages e.g. Portuguese (Specia et al., 2008), even though the tipology is different.

For the future, we should continue with this task by analysing other structures, improving the rules and their ordering, testing other methods (Woodsend and Lapata, 2011) (Siddharthan, 2011) using our dependency-based parsers (Aranzabe, 2008) (Bengoetxea et al., 2011), adapting the rules according to target audience etc.

9. Acknowledgements

Itziar Gonzalez-Dios's work is funded by a PhD grant from the Basque Government. This research was supported by the the Basque Government (IT344-10), and the Spanish Ministry of Science and Innovation (MICINN, TIN2010-202181).

10. References

I. Aduriz, E. Agirre, I. Aldezabal, I. Alegria, O. Ansa, X. Arregi, J.M. Arriola, X. Artola, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Maritxalar, M. Oronoz, K. Sarasola, A. Soroa, R. Urizar, and M. Urkia. 1998. A framework for the automatic processing of basque. In *Proceedings of Workshop on Lexical Resources for Minority Languages. First LREC Conference. Granada. 1998.".*

I. Aduriz, Aldezabal. I., I. Alegria, J.M. Arriola, X. Artola, A. Díaz de Ilarraza, N. Ezeiza, and Gojenola. 2003. Finite State Applications for Basque. In *EACL'2003 Workshop on Finite-State Methods in Natural Language Processing. pp. 3- 11".*

I. Aduriz, M. Aranzabe, J.M. Arriola, A. Atutxa, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa, and R. Urizar, 2006a. *Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing*.

I. Aduriz, K. Ceberio, and A. Díaz de Ilarraza. 2006b. Pronominal anaphora in basque: annotation of a real corpus. In *XXII Congreso de la SEPLN (Sociedad Espanola para el Procesamiento del Lenguaje Natural), pp. 99-104, ISSN: 1135-5948".*

E. Agirre, I. Alegria, X. Arregi, X. Artola, A. Díaz de Ilarraza, M. Maritxalar, K. Sarasola, and M. Urkia. 1992. Xuxen: A spelling checker/corrector for basque based in two-level morphology. In *Proceedings of NAACL-ANLP'92, 119-125. Povo Trento. 1992.*

E. Agirre, I. Aldezabal, J. Etxeberria, M. Iruskieta, E. Izagirre, K. Mendizabal, and E. Pociello. 2006. A methodology for the joint development of the basque wordnet and semcor. In *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC). ISBN 2-9517408-2-4. Genoa (Italy).".*

A.A. Al-Subaihin and H.S. Al-Khalifa. 2011. Al-baseet: A proposed simplification authoring tool for the arabic language. In *Communications and Information Technology (ICCIT), 2011 International Conference on*.

A. Alcázar. 2005. Towards linguistically searchable text. In *Proceedings of BIDE Summer School of Linguistics*.

I. Aldezabal, M. Aranzabe, A. Díaz de Ilarraza, A. Estarrona, K. Fernandez, and L. Uria. 2010. EPEC-RS: EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) rol semantikoekin etiketatzeko eskuliburua. Technical report.

I. Alegria, N. Ezeiza, I. Fernandez, and R. Urizar. 2003. Named entity recognition and classification for texts in basque. In *II Jornadas de Tratamiento y Recuperacin de Informacin, JOTRI, Madrid. 2003. ISBN 84-89315-33-7".*

S. M. Aluísio, L. Specia, T. A.S. Pardo, E. G. Maziero, and R. P.M. Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineer-*

ing, DocEng '08, pages 240–248, New York, NY, USA. ACM.

M. Aranzabe. 2008. Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala. In *Euskal Filologia Saila (UPV/EHU). EHUko Gipuzkoako Campuseko Joxe Mari Korta I+G+b zentroan, 2008ko urriaren 30ean*.

N. Areta, A. Gurrutxaga, I. Leturia, I. Alegria, X. Artola, A. Díaz de Ilarraza, N. Ezeiza, and A. Sologaistoa. 2007. ZT Corpus: Annotation and tools for Basque corpora. In *Corpus Linguistics Conference. Birmingham.*

B. Arrieta. 2010. Azaleko sintaxiaren tratamendua ikasketa automatikoko tekniken bidez: euskarako kateen eta perpausen identifikazioa eta bere erabilera komazuzentzaile batean. In *Informatika Fakultatea (UPV-EHU)*.

K. Bengoetxea, A. Casillas, and K. Gojenola. 2011. Testing the Effect of Morphological Disambiguation in Dependency Parsing of Basque. In *International Conference on Parsing Technologies (IWPT). 2nd Workshop on Statistical Parsing Morphologically Rich Languages (SPMRL)*.

J. Burstein. 2009. Opportunities for Natural Language Processing Research in Education. In *Computational Linguistics and Intelligent Text Processing*. Springer Berlin / Heidelberg.

A. Candido, Jr., E. Maziero, C. Gasperin, T. A. S. Pardo, L. Specia, and S. M. Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 34–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Carroll, G. Minnen, D. Pearce, Y. Canning, S. Devlin, and J. Tait. 1999. Simplifying Text for Language-Impaired Readers. volume 9th Conference of the European Chapter of the Association for Computational Linguistics.

D. Castro-Castro, R. Lannes-Losada, M. Maritxalar, I. Niebla, C. Pérez-Marqués, N.C. Alamo-Suarez, and A. Pons-Porrata. 2008. A multilingual application for automated essay scoring. In *Lecture Notes in Advances in Artificial Intelligence - LNAI 5290 - IBERAMIA ISBN 3-540-99308-8 Springer New York pp. 243-251"*.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*, COLING '96, pages 1041–1044, Stroudsburg, PA, USA. Association for Computational Linguistics.

N. Ezeiza. 2002. *CORPUSAK USTIATZEKO TRESNA LINGUISTIKOAK. Euskararen etiketatzaile morfosintaktiko sendo eta malgua.* Ph.D. thesis.

I. Gonzalez-Dios and M.J. Aranzabe. 2011. Euskarazko egitura sintaktikoen azterketa testuen sinplifikazio automatikorako: Aposizioak, erlatibozko perpausak eta denborazko perpausak. Master's thesis, University of Basque Country, September.

M. Iruskieta, A. Díaz de Ilarraza, and M. Lersundi. 2011. Unidad discursiva y relaciones retricas: un estudio acerca de las unidades de discurso en el etiquetado de un corpus en euskera. In *Procesamiento del Lenguaje Natural 47"*.

F. Karlsson, A. Voutilainen, J. Heikkila, and A. Anttila. 1995. *Constraint Grammar, A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

G. Labaka. 2010. EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language. Its use in SMT-RBMT-EBMT hybridation. In *Lengoaia eta Sistema Informatikoak Saila (UPV-EHU). Donostia. 2010ko martxoaren 29a*.

I. Laka. 1996. A brief grammar of euskara, the basque language.

A. Max. 2005. Simplification interactive pour la production de textes adapts aux personnes souffrant de troubles de la comprhension. volume Proceedings of Traitement Automatique des Langues Naturelles (TALN).

B. Oiarzabal, 2003. *A Grammar of Basque*, chapter Relatives. Mouton de Gruyter.

S. E. Petersen and M. Ostendorf. 2007. Text Simplification for Language Learners: A Corpus Analysis. *Electrical Engineering*, pages 69–72.

J. Rybing, C. Smith, and A. Sivervarg. 2010. Towards a Rule Based System for Automatic Simplification of texts. In *The Third Swedish Language Technology Conference (SLTC 2010)*.

H. Saggion, E. Gómez-Martínez, E. Etayo, A. Anula, and L. Bourg. 2011. Text Simplification in Simplext: Making Text More Accessible. *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*.

A. Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109.

A. Siddharthan. 2011. Text Simplification using Typed Dependencies: A Comparision of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11, Nancy, France, September. Association for Computational Linguistics.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA'2006*, pages 223–231, Columbus, Ohio, June.

L. Specia, S.M. Aluisio, and T.A.S Pardo. 2008. Manual de Simplificao Sinttica para o Portugułs. Technical Report NILC-TR-08-06, So Carlos-SP.

K. Woodsend and M. Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation

Alejandro Mosquera, Elena Lloret, Paloma Moreda

University Of Alicante
DLSI. Ap.de Correos 99. E-03080 Alicante, Spain
amosquera@dlsi.ua.es, elloret@dlsi.ua.es, moreda@dlsi.ua.es

Abstract

The Web 2.0, through its different platforms, such as blogs, social networks, microblogs, or forums allows users to freely write content on the Internet, with the purpose to provide, share and use information. However, the non-standard features of the language used in Web 2.0 publications can make social media content less accessible than traditional texts. For this reason we propose TENOR, a multilingual lexical approach for normalising Web 2.0 texts. Given a noisy sentence either in Spanish or English, our aim is to transform it into its canonical form, so that it can be easily understood by any person or text simplification tools. Our experimental results show that TENOR is an adequate tool for this task, facilitating text simplification with current NLP tools when required and also making Web 2.0 texts more accessible to people unfamiliar with these text types.

Keywords: Accessibility, Normalisation, Web 2.0

1. Introduction

The Web 2.0, through its different platforms, such as blogs, social networks, microblogs, or forums allows users to freely write content on the Internet, with the purpose to provide, share and use information. It is known that this type of platforms are among the top visited websites[1], and their interest is growing more and more.

However, despite of the great potential of this user-generated content, it has several well-known drawbacks, concerning what is communicated and how it is communicated. On the one hand, the information users provide has not always the same level of reliability, and therefore wrong or inaccurate information can be considered as correct one (Scanfeld et al., 2010), (Mendoza et al., 2010). On the other hand, the Internet, and in particular, the Web 2.0, has an informal nature, since there is not any restriction regarding the language employed for posting on-line information. To name just a few: i) the use of emoticons (e.g., :-P); ii) non-standard abbreviations (e.g., LOL – *laugh out loud*) and contractions (e.g., abt – *about*); iii) frequent typos and spelling errors (e.g., *lasi*, instead of lazy); and iv) a lot of use of interjections and letter-repetitions (e.g., *yeaahhhhhh!*).

These non-standard features can make Web 2.0 publications less accessible than traditional texts to people unfamiliar with this type of lexical variants or people with disabilities. To date and to our knowledge, studies on text accessibility focus on simplification strategies. For this reason, performing a normalisation process is a step prior to simplification for non-accessible Web 2.0 texts.

Therefore, the objective of this paper is to suggest TENOR, a lexical approach for normalising Web 2.0 texts. Given a noisy sentence either in Spanish or English, our aim is to transform it into its canonical form, so that it can be easily understood by any person or text simplification tools. By achieving this goal, texts could be transcribed using standard and common language, making them easier and more comprehensible, and thus facilitating straightforward the reading comprehension process for people with difficulties, as well as the use of existing automatic tools for carrying out other tasks, such as text simplification or summarisation.

This article is organised as follows. In Section 2, the state of the art is reviewed, discussing existing research works dealing with text simplification and normalisation, and stressing the differences of our approach with respect to them. Further on, Section 3 describes our normalisation approach for very informal texts. Next, in Section 4 we described the evaluation conducted, together with a in-depth discussion of the results obtained, and finally Section 5 concludes this paper and outlines future work.

2. Related Work

In the recent years, making information more accessible to everybody is a relevant issue which is gaining a lot of attention among the research community. One of the research areas devoted to this purpose is Text Simplification whose aim is to rewrite the information into a simpler way in order to help users to comprehend the information that, if left unedited, would be too complex to understand. To this end, the types of simplification include: i) lexical, which substitutes non-frequent words to more common ones (Biran et al., 2011); ii) syntactic, which splits difficult and large sentences into simpler ones (Evans, 2011); and iii) semantic, which attempts to provide definitions for difficult expressions and/or non-literal meaning (Barnden, 2008). Initiatives such as Simple Wikipedia[2], Noticias Fácil[3] as well as several past and on-going projects, as for instance, SkillSum (Williams and Reiter, 2008), Simplext (Saggion et al., 2011), or FIRST[4], constitute good contexts for mak-

[1] http://www.alexa.com/topsites

[2] http://simple.wikipedia.org/wiki/Main_Page

[3] http://www.noticiasfacil.es/ES/Paginas/index.aspx

[4] http://www.first-asd.eu/

ing progress within this area, thus being beneficial for individuals with low literacy (Candido et al., 2009), physical and cognitive disabilities (Daelemans et al., 2004),(Huenerfauth et al., 2009), or even language learners (Petersen and Ostendorf, 2007).

However, as these systems are designed to work with standard texts, the special features of the language used in the Web 2.0 can difficult their processing.

Furthermore, another subfield of Natural Language Processing (NLP) deals with Text Normalisation of user-generated content.

The process of text normalisation basically cleans an input word or sentence by transforming all non-standard lexical or syntactic variations into their canonical forms. From the existing literature, we have identified three major trends to tackle this task. The first one relies on machine translation techniques (Aw et al., 2006), (López et al., 2010) the second focuses on orthographic correction approaches (Liu et al., 2011), and the third one takes as a basis a combination of lexical and phonetic edit distances (Han and Baldwin, 2011), (Gouws et al., 2011). Among them, we would like to outline the research works proposed in (Han and Baldwin, 2011) and (Liu et al., 2011). In the former, supervised classification techniques are employed for identifying ill-formed words, which are then normalised by extracting the best candidate among several ones, using a set of rules. In the latter, a letter transformation approach is proposed through the use of the noisy channel model (Shannon, 1948).

To the best of our knowledge, none of the previous works have used a multilingual strategy, thus being restricted to the English language only. Therefore, this paper proposes the use of TENOR, a multilingual normalisation tool for the Web 2.0 with the purpose of obtaining the canonical form of a text, so it can be more accessible to more people and for current NLP simplification tools.

In the next section, our approach will be explained in detail.

3. TENOR, TExt NORmalisation Approach

In this section we explain TENOR, our text normalisation approach based on a combination of lexical and phonetic edit distances for short English and Spanish texts belonging to the Web 2.0.

Our normalisation process comprises two steps: First, it uses a classification method to detect non-standard lexical variants or words out of vocabulary. Second, the selected words in the previous step are replaced to their original standard form. Each of this stages are going to be explained in more detail.

3.1. Out-of-Vocabulary detection

In this section we refer to words outside the vocabulary as those that are not part of standard English or Spanish vocabulary and need to be standardised. However, the detection of such words is not a trivial task: The presence of proper names, cities, neologisms and acronyms, as well as the richness of the language makes it difficult to know when a word belongs to the language or otherwise is a lexical variant (see Table 1).

	OOV word	Canonic word
a)	sucess	success
b)	rite	right
c)	playin	playing
d)	emocion	emoción
e)	mimir	dormir
f)	separa2	separados

Table 1: Out of vocabulary and canonic pairs examples from Web 2.0 texts. Examples from *a* to *c* correspond to English and the ones from *d* to *f* to Spanish.

In TENOR, OOV words are detected with a dictionary lookup. In order to do this, we use custom-made lexicons built over the expanded English and Spanish Aspell[5] dictionaries. These are augmented with domain-specific knowledge such as the Spell Checking Oriented Word Lists (SCOWL)[6] for English, and country names, cities, acronyms and common proper names[7] for Spanish. Heuristics based on capitalisation of words are employed to identify named entities and acronyms. Likewise, some special Twitter tags are used to perform a slight syntactic disambiguation, such as: @(User Name) # (Tag), RT (Retweet) and TT (Trending Topic), thus avoiding the processing of such elements.

3.2. Substitution of Lexical Variants

This section discusses the different steps carried out to replace the words classified as OOV with their normalised form. In order to do this, several substages are proposed. First, in Section 3.2.1 the filtering techniques employed to "clean" texts are introduced. In Section 3.2.2 we detail the process of replacing common word transformations. Then, in Section 3.2.3 the use of phonetic indexing in order to obtain lists of words with equivalent pronunciations by building a phone lattice is described. Subsequently, in Section 3.2.4 we explain how this lattice is used in order to identify possible candidates to replace the non-normative lexical variants. Finally, in Section 3.2.5 we show how the use of language models can help to select the most appropriate canonical word from the list of phonetic candidates.

3.2.1. Filtering

First,all non-printable characters and non-standard punctuations with the exception of emoticons are eliminated using regular expressions. While these may be beyond the scope of the study and therefore not to be considered lexical variants, their filtering would negatively impact another NLP tools such as opinion mining or sentiment analysis.

3.2.2. Common Word Tranformations

The second step of the analysis is to identify common word transformations such as abbreviations and transliterations, which are replaced by their equivalent standard form: i) Word-lengthening compression (see Table 3, example c) is

[5]http://aspell.net

[6]http://wordlist.sourceforge.net/

[7]http://es.wikipedia.org

performed by applying heuristic rules to reduce the repetition of vowels or consonants within a word *(nooo! - no!, goooooolll - gol)*; ii) There are numbers whose pronunciation is often used to shorten the length of the message *(ning1 - ninguno)* or combination of letters and *(h0us3 - house)*. In these cases they were replaced by following a transliteration conversion table. In Table 2, each number is assigned its most frequent meanings when it appears as a part of a word; iii) Emoticon translation (see Table 3, example b) was made by grouping smileys into two categories (happy, sad), thus being replaced by their textual equivalent using simple heuristic rules based on regular expressions; iv) Simple case restoration techniques were applied to wrong-cased words *(GrEaT - great)*.

Nº	English	Spanish
0	0, zero, o	0, cero, o
1	1, one	1, uno
2	2, two, too	2, dos
3	3, three, e	3, tres, e
4	4, for, a	4, cuatro, a
5	5, five, s	5, cinco, s
6	6, six, g	6, seis, g
7	7, seven, t	7, siete, t
8	8, eight	8, ocho
9	9, nine, g	9, nueve, g

Table 2: Common numeric transliterations found in Web 2.0 English and Spanish texts.

3.2.3. Phonetic Indexing

The aim of this stage is to obtain a list of candidate terms for each OOV words detected in previous stages. In order to do this, TENOR obtains lists of words with equivalent pronunciations using phonetic indexing techniques to build a phone lattice. OOV words are matched against this phone lattice with the metaphone algorithm (Philips, 2000) to obtain such list of substitution candidates. The metaphone algorithm allows to represent the pronunciation of a word using a set of rules. In particular the double-metaphone reference implementation for English and an adaptation of the metaphone for the Spanish language[8]. For example, the Spanish metaphone *(JNTS)* can index the words *gentes, jinetas, jinetes, juanetes, juntas, juntos* between others and the English metaphone *(PRXS)* can index the words *purses, prices, precise, praises* among others.

Moreover, there are acronyms and abbreviated forms that can not be detected properly with phonetic indexing techniques *(lol - laugh out loud)*. For this reason, TENOR uses an exception dictionary manually built upon an equivalence table with 46 of the most common Spanish abbreviations *(qtal - qué tal), (xfa - por favor)* and 196 English Internet abbreviations and slang words[9] that need special treatment because their low similarity with their equivalent standard form *(gotta - going to), (omg - oh my god)*.

[8]http://github.com/amsqr/Spanish-Metaphone

[9]http://en.wiktionary.org/wiki/Appendix:English_internet_slang

3.2.4. Lexical Similarity

Once the possible candidates associated to a OOV word are obtained, the lexical similarity between each candidate and the OOV word is computed. For this, we use the Gestalt pattern matching algorithm (Ratcliff and Metzener, 1988). This algorithm provides a string similarity score based on the maximum common subsequence principle between 0 and 100, where 0 is minimum similarity and 100 is maximum similarity. This score is calculated between the OOV word and its candidate list, empirically discarding candidates with similarity values lower than 60.

3.2.5. Candidate Selection

In order to obtain the final substitution candidate when there are more than one candidate word with the same similarity value a trigram language model has been used. TENOR contains 2 models both for English and Spanish texts, trained with the Brown corpus (Kucera and Francis, 1967) and the CESS-ESP (Martí and Taulé, 2007) respectively, with smoothing techniques (Chen and Goodman, 1996). This task has been implemented with the NLTK NgramModel class (Bird, 2006) for determining the replacement that minimises the perplexity, taking the latter as a measure of model quality.

4. Evaluation and Results

This section describes the evaluation process and the analysis of the results obtained with TENOR. First, the used corpora is introduced in Section 4.1. Subsequently, TENOR evaluation is explained in Section 4.2. Finally, the obtained results are discussed in Section 4.3.

4.1. Corpus

Two different corpora extracted from Twitter have been used in the evaluation process. Twitter[10] is an on-line microblogging service that enables its users to send and read textual messages of up to 140 characters. Due to this space constrain and its informal nature it can be considered a good source of short and noisy texts. Han's Twitter dataset[11] has been used for English texts and, following the same tagging scheme, a hand-annotated corpus of 1000 Tweets texts has been used for Spanish results[12]. In both cases, tagged words are annotated as out of vocabulary (OOV), inside the vocabulary (IV) or non-processable (NO). Also, for each OOV word its canonic version is provided.

4.2. Evaluation

We have evaluated TENOR performance in terms of precision and recall (Tang et al., 2005) taking into account OOV detection and normalisation separately (see Table 4). The obtained results were matched against the gold standard described in 4.1.

4.3. Results

TENOR results improve state-of-the-art approaches, with a 92% and a 82% F1 in OOV detection and OOV normalisation respectively (see Table 5).

[10]http://www.twitter.com

[11]http://www.csse.unimelb.edu.au/research/lt/resources/lexnorm/

[12]http://gplsi.dlsi.ua.es/gplsi11/content/twitter-norm-dataset

	Raw Spanish	Normalised Spanish
a)	tdo StO no s cierT, stams caNsa2	todo esto no es cierto, estamos cansados
b)	xfa apoyo xa 1 niño d 3 añits	por favor apoyo para 1 niño de 3 añits
c)	mal momemto para sufrur!	mal momento para sufrir!
d)	bamos a x ellos nesecitamos el apollo!!!!!	vamos a por ellos necesitamos el apoyo!
e)	amunt! valencia, visca el barça!	aumento! Valencia, busca el F.C. Barcelona!
f)	el no aprobara	el no aprobara
	Raw English	**Normalised English**
g)	whn ur talking to some1 an u say tht	When you are talking to someone and you say that.
h)	Talkin abt this wee lasi cawd sophie:(	Talking about this week lazy caw sophie I'm sad.
i)	WAAAAAAAy up great!	Way up great!
j)	its my last wish to see u plz	Its my last wish to see you please.

Table 3: Raw and normalised pairs of Spanish and English Web 2.0 examples.

	(OOV)	(IV)
Found	A	B
Not Found	C	D

Precision:	P=A/(A+B)
Recall:	R=A/(A+C)
F1:	F=2PR/(P+R)

Table 4: Evaluation measures used in this study.

Task	Precision	Recall	F1
TENOR Eng. OOV	**91.7%**	**95.2%**	**93.4%**
TENOR Sp. OOV	**82.7%**	**98%**	**89.7%**
Han-Baldwin2011 OOV	61.1%	85.3%	71.2%
Han-Baldwin2011	75.3%	**75.3%**	75.3%
TENOR Eng.	88.9%	55.3%	68.2%
TENOR Eng. w/except.	**91.2%**	74.5%	**82.1%**
TENOR Sp.	94.1%	56%	70.2%
TENOR Sp. w/except.	**96.1%**	73%	**83%**

Table 5: Evaluation of out of vocabulary detection and normalisation results with and without the exception dictionary for English and Spanish Twitter texts.

Taking into account the obtained results, the use of the exception dictionary significantly enhances the normalisation of both English and Spanish texts. It can be noticed that Spanish normalisation results are higher, although its dictionary contains less entries than the English one. This is directly related to the results obtained in the OOV detection, in which the Spanish version of TENOR obtained slightly lower results (Fmeasure) than its English version. We can conclude that in Spanish is more difficult to detect OOV words and this is because there is a greater number of exceptions. By using the exception dictionary for English results also were improved, but this was expected since the exception dictionary was of greater size.

The obtained results show the capability of TENOR as a tool for improving Web 2.0 texts accessibility by facilitating the work of current NLP simplification tools. Moreover, it also makes user-generated content more accessible to people unfamiliar with these text types.

5. Conclusions and Future Work

In this paper we have presented TENOR, a multilingual text normalisation approach for Web 2.0 texts. We have demonstrated that with TENOR noisy and difficult to understand English and Spanish texts can be converted into their canonic form. The substitution of non-normative vocabulary present in Web 2.0 texts, results in texts easier to understand and therefore makes the Web 2.0 new textual genres more accessible to everybody. This is a first step to facilitate the understanding of texts by easing the access to information using the new forms of communication available in the Web 2.0.

In the future we plan to extend TENOR by adding support to additional languages. Moreover, as a long term goal we would like to integrate this tool with text simplification strategies.

Acknowledgements

This paper has been partially supported by Ministerio de Ciencia e Innovación - Spanish Government (grant no. TIN2009-13391-C04-01), Conselleria d'Educación - Generalitat Valenciana (grant no. PROMETEO/2009/119, ACOMP/2010/286 and ACOMP/2011/001) and the European Commission under the Seventh (FP7 - 2007-2013) Framework Programme for Research and Technological Development through the FIRST project (FP7-287607). This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

6. References

Aiti Aw, Min Zhang, Juan Xiao, and Jian Su. 2006. A phrase-based statistical model for sms text normalization. *Proceedings of the COLING/ACL*, pages 33–40.

John Barnden. 2008. Challenges in natural language processing: the case of metaphor (commentary). *International Journal of Speech Technology*, 11:121–123.

Or Biran, Samuel Brody, and Noemie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501, Portland, Oregon, USA, June. Association for Computational Linguistics.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

Arnaldo Candido, Jr., Erick Maziero, Caroline Gasperin, Thiago A. S. Pardo, Lucia Specia, and Sandra M. Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, EdAppsNLP '09, pages 34–42, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL-1996)*, pages 310–318.

Walter Daelemans, Anja Höthker, and Erik Tjong Kim Sang. 2004. Automatic sentence simplification for subtitling in dutch and english. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1045–1048.

Richard J. Evans. 2011. Comparing methods for the syntactic simplification of sentences in information extraction. *Literacy and Linguist Computing*, 26(4):371–388.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual Bearing on Linguistic Variation in Social Media. *ACL Workshop on Language in Social Media (LSM)*.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 368–378, Portland, Oregon, USA, June. Association for Computational Linguistics.

Matt Huenerfauth, Lijun Feng, and Noemie Elhadad, 2009. *Comparing evaluation techniques for text readability software for adults with intellectual disabilities*, pages 3–10. ACM.

Henry Kucera and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, USA.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 71–76, Portland, Oregon, USA, June. Association for Computational Linguistics.

Veronica López, Rubén San-Segundo, Roberto Martín, Julian David Echeverry, and Syaheera Lutfi. 2010. Sistema de traducción de lenguaje SMS a castellano. In *XX Jornadas Telecom I+D*, Valladolid, Spain, September.

Maria Antonia Martí and Mariona Taulé. 2007. Cess-ece: corpus anotados del español y catalán. *Arena Romanistica. A new Nordic journal of Romance studies*, 1.

Marcelo Mendoza, Barbara Poblete, and Carlos Castillo, 2010. *Twitter Under Crisis: Can we trust what we RT?*, volume 1060, page 10. ACM Press.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners : A corpus analysis. *Electrical Engineering*, (SLaTE):69–72.

Lawrence Philips. 2000. The double metaphone search algorithm. *C/C++ Users Journal*, 18:38–43, June.

John W. Ratcliff and David E. Metzener. 1988. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46–72, July.

Horacio Saggion, Elena Gómez-Martínez, Alberto Anula, and Esteban Bourg, Lorena an dEtayo. 2011. Text simplification in simplext: Making texts more accessible. *Procesamiento del Lenguaje Natural*, 47:341–342.

Daniel Scanfeld, Vanessa Scanfeld, and Elaine L Larson. 2010. Dissemination of health information through social networks: twitter and antibiotics. *American Journal of Infection Control*, 38(3):182–188.

Claude. E. Shannon. 1948. A mathematical theory of communication. *The Bell Systems Technical Journal*, 27:379–423.

Jie Tang, Hang Li, Yunbo Cao, and Zhaohui Tang. 2005. Email data cleaning. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 489–498, New York, NY, USA. ACM Press.

Sandra Williams and Ehud Reiter. 2008. Generating basic skills reports for low-skilled readers*. *Nat. Lang. Eng.*, 14:495–525, October.

What Can Readability Measures Really Tell Us About Text Complexity?

Sanja Štajner, Richard Evans, Constantin Orăsan, and Ruslan Mitkov

Reseach Institute in Information and Language Processing
University of Wolverhampton
S.Stajner@wlv.ac.uk, R.J.Evans@wlv.ac.uk, C.Orasan@wlv.ac.uk, R.Mitkov@wlv.ac.uk

Abstract

This study presents the results of an initial phase of a project seeking to convert texts into a more accessible form for people with autism spectrum disorders by means of text simplification technologies. Random samples of Simple Wikipedia articles are compared with texts from News, Health, and Fiction genres using four standard readability indices (Kincaid, Flesch, Fog and SMOG) and sixteen linguistically motivated features. The comparison of readability indices across the four genres indicated that the Fiction genre was relatively easy whereas the News genre was relatively difficult to read. The correlation of four readability indices was measured, revealing that they are almost perfectly linearly correlated and that this correlation is not genre dependent. The correlation of the sixteen linguistic features to the readability indices was also measured. The results of these experiments indicate that some of the linguistic features are well correlated with the readability measures and that these correlations are genre dependent. The maximum correlation was observed for fiction.

Keywords: text simplification, readability, autism spectrum disorders

1. Introduction

Text simplification can be regarded as the process of converting input text into a more accessible form. The conversion process may be facilitated by research in various areas of NLP, including lexical simplification (Yatskar et al., 2010), anaphora resolution (Mitkov, 2002), word sense disambiguation (Escudero et al., 2000), syntactic simplification (Siddharthan, 2006; Evans, 2011), text summarisation (Orăsan and Hasler, 2007), or image retrieval (Bosma, 2005).

In the context of personalisable applications, it is necessary for systems not only to simplify text, but also to discriminate between material that should be simplified and material that should not be, for the benefit of a particular user. This discrimination can be realised by quantifying the difficulty of the material by means of various features of the text, and comparing those feature values with thresholds specified in user preferences.

The work described in this paper is part of an ongoing project that develops tools to help readers with autism spectrum disorders (ASD). One of the prerequisites for this research is to have a way to assess the difficulty of texts. A set of metrics is proposed with the aim of quantifying the difficulty of input documents with respect to their requirements. This set contains readability indices and metrics inspired by the needs of people with ASD. Documents from several genres are evaluated with regard to these metrics and the correlation between them is reported.

1.1. Requirements of Users with Autism Spectrum Disorders

This paper presents research undertaken in the initial phase of FIRST,[1] a project to develop language technology (LT) that will convert documents from various genres in Bulgarian, English, and Spanish into a more accessible form for readers with autism spectrum disorders (ASD).

ASD are defined as neurodevelopmental disorders characterised by qualitative impairment in communication and stereotyped repetitive behaviour. They are serious disabilities affecting approximately 60 people out of every 10 000 in the EU. People with ASD usually have language deficits with a life-long impact on their psychosocial functioning. These deficits are in the comprehension of speech and writing, including misinterpretation of figurative language and difficulty understanding complex instructions (Minshew and Goldstein, 1998). In many cases, people with ASD are unable to derive the gist of written documents (Nation et al., 2006; O'Connor and Klein, 2004; Frith and Snowling, 1983).

Written documents pose various obstacles to reading comprehension for readers with ASD. These include:

1. Ambiguity in meaning:
   (a) Figurative language such as metaphor and idioms,
   (b) Non-literal language such as sarcasm,
   (c) Semantically ambiguous words and phrases,
   (d) Highly specialised/technical words and phrases.
2. Structural complexity:
   (a) Morphologically, orthographically, and phonetically complex words,
   (b) Syntactically complex sentences,
   (c) Inconsistent document formatting.

A detailed study of user requirements derived from a focus group partially supported the initial hypothesis of their reading comprehension difficulties. The focus group made recommendations for the automatic simplification

[1] A Flexible Interactive Reading Support Tool (http://www.first-asd.eu).

of phenomena at various linguistic levels. This includes the automatic expansion and elaboration of acronyms and abbreviations (obstacle 1d); the replacement of ambiguous words/homographs by less ambiguous words (obstacle 1c); the substitution of anaphoric references by their antecedents, especially in the case of zero anaphora (obstacle 1c); the rewriting of long sentences as sequences of short sentences, the conversion of passive sentences into active sentences (obstacle 2b); and the translation of phraseological units such as collocations, idioms, and ironic/sarcastic statements into a more literal form (obstacles 1a and 1b).
In addition to the removal of obstacles to reading comprehension, recommendations were also made for the addition of indicative summaries, multimedia, and visual aids to the converted documents output by FIRST.

1.2. Readability Indices

Independent of the specific requirements of readers with ASD, readability indices are one means by which the reading difficulty of a document can be estimated. DuBay (2004) notes that over 200 readability formulae have been developed so far, with over 1 000 studies of their application published. In the research described in the present paper, the Flesch Reading Ease score (Flesch, 1949), the Kincaid readability formula (Kincaid et al., 1986), the Fog Index (Gunning, 1952), and SMOG grading (McLaughlin, 1969) metrics were selected for this purpose. Considering each in turn:
The Flesch Reading Ease score is obtained by the formula:

$$Score = 206.835 - (1.015 \times ASL) - (84.6 \times ASW)$$

Here, ASL denotes the average sentence length and ASW the average number of syllables per word. The Flesch Reading Ease Formula returns a number from 1 to 100, rather than grade level. Documents with a Flesch Reading Ease score of 30 are considered "very difficult" while those with a score of 70 are considered "easy" to read. The software developed in FIRST is therefore required to convert documents into a form with a Reading Ease Score higher than 90, commensurate with fifth grade reading level.
The Flesch-Kincaid readability formula[2] is a simplified version of the Flesch Reading Ease score. It is based on identification of the average sentence length of the document to be assessed (ASL) and the average number of syllables per word in the document (ASW). The formula estimates readability by US grade level (GL):

$$GL = (0.4 \times ASL) + (12 \times ASW) - 15$$

The Fog Index (Gunning, 1952) exploits two variables: average sentence length and the number of words containing more than two syllables ("$hard\ words$") for each 100 words of a document. This index returns the US Grade Level (GL) of the input document, according to the formula:

$$GL = 0.4 \times (average\ sentence\ length + hard\ words).$$

The SMOG grading (McLaughlin, 1969) is computed by considering the polysyllable count, equivalent to the number of words that contain more than two syllables in 30 sentences, and applying the following formula:

$$SMOG\ grading = 3 + \sqrt{polysyllable\ count}$$

It has been noted that the SMOG formula is quite widely used, particularly in the preparation of US healthcare documents intended for the general public.[3]
The selection of these standard readability metrics was made due to the observation that, although based on different types of information, they all demonstrate significant correlation in their prediction of the relative difficulty of the collections of documents assessed in the research described in this paper.
The standard readability metrics were computed using the GNU *style* package, which exploits an automatic method for syllable identification. Manual studies of the efficacy of this module suggest that it performs with an accuracy of roughly 90%, similar to state of the art part-of-speech taggers.

2. Related Work

Previous research has shown that the average US citizen reads at the seventh grade level (NCES, 1993). Experts in health literacy have recommended that materials to be read by the general population should be written at fifth or sixth grade level (Doak et al., 1996; Weiss and Coyne, 1997). The FIRST project aims to produce documents suitable for users with reading comprehension problems. Due to the reading difficulties of people with ASD, documents output by the software developed in the project should not exceed the fifth grade level (suitable for people with no reading comprehension difficulties at ten or eleven years old). Together, these constraints emphasise the desirability of consistent and reliable methods to quantify the readability of documents.
In Flesch (1949), it was found that documents presenting fictional stories lay in the range $70 \leq Score \leq 90$. Only comics were assigned a higher score for reading ease than this. The most difficult type of document was that of scientific literature, with $0 \leq Score \leq 30$. During the 1940s, the Reading Ease Scores of news articles were at the sixteenth grade level. It is estimated that in contemporary times, this has been reduced to eleventh grade level.
The set of linguistic features employed in the research described in this paper (Section 3.2.) shares some similarity with the variables shown by Gray and Leary (1935) to be

[2] To avoid confusion, in the current paper, the *Flesch-Kincaid readability formula* will hereafter be referred to as the *Kincaid readability formula*.

[3] For example, the Harvard School of Public Health provides guidance to its staff on the preparation of documents for access by senior citizens that is based on the SMOG formula (http://www.hsph.harvard.edu/healthliteracy/files/howtosmog.pdf, last accessed 1st March 2012).

closely correlated with reading difficulty. These variables include the number of first, second, and third person pronouns (correlation of 0.48), the number of simple sentences within the document (0.39), and the number of prepositional phrases occurring in the document (0.35). There is also some similarity with features exploited by Coleman (1965) in several readability formulae. These features include counts of the numbers of pronouns and prepositions occurring in each 100 words of an input document.

DuBay (2004) presents the arguments of several researchers who criticise the standard readability indices on numerous grounds. For example, the metrics have been noted to disagree in their assessment of documents (Kern, 2004). However, DuBay defends their use, arguing that the important issue is the degree of consistency that each formula offers in its predictions of the difficulty of a range of texts and the closeness with which the formulae are correlated with reading comprehension test results.

Research by Coleman (1971) and Bormuth (Bormuth, 1966) highlighted a close correlation between standard readability metrics and the variables shown to be indicative of reading difficulty. These findings motivate the current investigation into potential correlation between standard readability metrics and the metrics sensitive to the occurrence of linguistic phenomena.

3. Methodology

This section describes the methodology employed in order to explore potential correlations between the standard readability indices and the linguistic features used to measure the accessibility of different types of document for readers with ASD. It contains a description of the corpora (Section 3.1.), details of the linguistic features of accessibility that are pertinent for these readers (Section 3.2.), and details of the means by which the values of these features were automatically obtained (Section 3.3.).

3.1. Corpora

The LT developed in the FIRST project is intended to convert Bulgarian, English, and Spanish documents from fiction, news, and health genres[4] into a form facilitating the reading comprehension of users with ASD. The current paper focuses on the processing of documents written in English.

Collections of documents from these genres were compiled on the recommendation of clinical experts within the project consortium. This recommendation was based on the prediction that access to documents of these types would both motivate research into the removal of a broad spectrum of obstacles to reading comprehension and also serve to improve perceptions of inclusion on the part of readers with ASD. In the current paper, the assessment of readability is made with respect to the following document collections (Table 1):

1. **NEWS** - a collection comprising reports on court cases in the METER corpus (Gaizauskas et al., 2001) and articles from the PRESS category of the FLOB corpus.[5] The documents selected from FLOB were each of approximately 2 000 words in length. The news articles from the METER corpus were rather short; none of them had more than 1 000 words. We included only documents with at least 500 words;
2. **HEALTH** - a collection comprising healthcare information contained in a collection of leaflets for distribution to the general public, from categories *A01*, *A0J*, *B1M*, *BN7*, *CJ9*, and *EDB* of the British National Corpus (Burnard, 1995). This sample contains documents with considerable variation in word length;
3. **FICTION** - a collection of documents from the FICTION category of the FLOB corpus. Each is approximately 2 000 words in size; and
4. **SIMPLEWIKI** - a random selection of simplified **encyclopaedic** documents, each consisting of more than 1 000 words, from Simple Wikipedia.[6] This collection is included as a potential model of accessibility. One of the goals of the research described in this paper is to compare the readability of other types of document from this "standard".

Corpus	Words	Texts
SimpleWiki	272,445	170
News	299,685	171
Health	113,269	91
Fiction	243,655	120

Table 1: Size of the corpora

3.2. Linguistic Features of Document Accessibility

The obstacles to reading comprehension faced by people with ASD when seeking to access written information were presented in Section 1.1. The features presented in this section are intended to indicate the occurrence of these obstacles in input documents. Thirteen features are proposed as a means of detecting the occurrence of the different types of obstacle to reading comprehension listed in Section 1.1. Related groups of features are presented below.

(1) **Features indicative of structural complexity**: This group of ten features was inspired by the syntactic concept of the projection principle (Chomsky, 1986) that "lexical structure must be represented categorically at every syntactic level". This implies that the number of noun phrases in a sentence is proportional to the number of nouns in that sentence, the number of verbs in a sentence is related to the number of clauses and verb phrases, etc. The values of nine of these features were obtained by processing the output of *Machinese Syntax*[7] to detect the

[4] In this paper, we use the term *health* to denote documents from the genre of education in the domain of health.

[5] Freiburg-LOB Corpus of British English (http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM)

[6] http://simple.wikipedia.org

[7] http://www.connexor.eu

Feature	Indicator of
Nouns (N)	References to concepts/entities
Adjectives (A)	Descriptive information about concepts/entities
Determiners (Det)	References to concepts that are not proper names, acronyms, or abbreviations
Adverbs (Adv)	Descriptive information associated with properties of and relations between concepts/entities
Verbs (V)	Properties of and relations between concepts/entities
Infinitive markers (INF)	Infinitive verbs (a measure of syntactic complexity)
Coordinating conjunctions (CC)	Coordinated phrases
Subordinating conjunctions (CS)	Subordinated phrases, including phrases embedded at multiple levels
Prepositions (Prep)	Prepositional phrases (a well-cited source of syntactic ambiguity and complexity)

Table 2: Features (structural complexity)

occurrence of words/lemmas with particular part-of-speech tags (Table 2). As the tenth feature we proposed *Sentence complexity* (Compl) in terms of number of verb chains. It was measured as the ratio of the number of sentences in the document containing at most one verb chain to the number containing two or more verb chains. To illustrate, the sentence:

> *I am consumed with curiosity, and I cannot rest until I know why this Major Henlow should have sent the Runners after you.*

contains four verb chains: {*am consumed*}, {*cannot rest*}, {*know*}, and {*should have sent*}. This feature exploits the functional tags assigned to different words by *Machinese Syntax* (Section 3.3.1.).

(2) **Features indicative of ambiguity in meaning**: This group of three features (Table 3) is intended to indicate the amount of semantic ambiguity in the input document.

Feature	Indicator of
Pronouns (Pron)	Anaphoric references
Definite descriptions (defNP)	Anaphoric references
Word senses (Senses)	Semantic ambiguity

Table 3: Features (ambiguity in meaning)

In all three cases, the difficulties caused by the feature arise as a result of doubts over the reference to concepts in the domain of discourse by different linguistic units (words and phrases). The values of these features are obtained by processing the output of *Machinese Syntax* to detect both the occurrence of words/lemmas with particular parts of speech and the functional dependencies holding between different words, and exploitation of WordNet as a source of information about the senses associated with content words in the input text.
These features were calculated as averages per sentence. The only exception was the feature *Senses* which was computed as the average number of senses per word.

3.3. Extraction of Linguistic Features

A user requirements analysis undertaken during the initial stage of the project motivated the development of features of accessibility based on the occurrence of various linguistic phenomena in an input document. Given that these are complex and difficult to detect automatically, the linguistic features are based on indicative morpho-syntactic information that can be obtained via existing NLP resources.
Derivation of the feature values depends on exploitation of two language technologies: Connexor's *Machinese Syntax* functional dependency parser (Tapanainen and Jarvinen, 1997) and the generic ontology, WordNet (Fellbaum, 1998). The detection process is based on the assumption that words with particular *morphological* and *surface syntactic* tags assigned by *Machinese Syntax* indicate the occurrence of different types of linguistic phenomenon.
One caveat that should be made with regard to the values obtained for these features is that they exploit language processing technology that is imperfect in its accuracy and coverage. The efficacy of Connexor's *Machinese Syntax*, used to obtain the values for the linguistic features, is described in (Tapanainen and Jarvinen, 1997).

3.3.1. Functional Dependencies

The values of two features, *defNP* and *Compl*, are obtained by reference to the functional dependencies detected by *Machinese Syntax* between words in the input documents.
The feature defNP is intended to obtain the number of definite noun phrases occurring in each sentence of an input document. This number is measured by counting the number of times that functional dependencies occur between tokens with the lemma *the*, *this*, and *that* and tokens with a nominal surface syntactic category.
The feature *Compl*, which relies on identification of the verb chains occurring in each sentence of a document (see Section 3.2.), exploits analyses provided by the parsing software. Verb chains are recognised as cases in which verbs are assigned either *finite main predicator* or *finite auxiliary predicator* functional tags by *Machinese Syntax*.

3.3.2. WordNet

Word sense ambiguity (*Senses*) was detected by exploitation of the WordNet ontology (Fellbaum, 1998). Input documents are first tokenised and each token disambiguated in terms of its surface syntactic category by *Machinese Syntax*. The number of concepts linked to the word when used with that category were then obtained from WordNet. The extraction method thus exploits some limited word sense disambiguation as a result of the operation of the parser. As noted earlier (Section 3.2.), the feature *Senses* was calculated as the average number

Corpus	Kincaid	Flesch	Fog	SMOG	ch/w	syl/w	w/s
SimpleWiki	**7.49**	**69.91**	**10.35**	**9.78**	**4.67**	**1.43**	**16.05**
News	**9.39**	**64.98**	**12.28**	**10.77**	4.66	1.43	**20.90**
Health	7.84	69.31	10.83	10.07	4.63	1.42	**17.13***
Fiction	**5.05**	**83.06**	**7.85**	**7.90**	**4.29**	**1.30**	**13.58**

Table 4: Readability indices and related features

of senses per word. Therefore, multiple occurrences of the same ambiguous word will increase the value of this feature.

4. Results

The study presented in this paper comprises three parts. In the first, a comparison is made between the values obtained for the four readability indices and the factors that they exploit (average numbers of characters and syllables per word, average number of words per sentence) in their assessment of the corpora (SimpleWiki, News, Health, and Fiction). If the intuitive assumption is valid, that SimpleWiki represents a corpus of simplified texts (a "gold standard"), then this comparison will indicate how far documents from the news, health, and fiction genres (important for the social inclusion of people with ASD) lie from this 'gold standard'.

In the second part, the use of thirteen linguistic features is explored. Ten of the linguistic features are based on the frequency of occurrence of surface syntactic tags, one is based on sentence complexity expressed in terms of the number of verb chains that they contain, another provides an approximation of the number of definite noun phrases used in the text, and the final feature measures the average level of semantic ambiguity of the words used in the text. The values obtained for these features for each of the corpora are compared.

In the third part of the study, potential correlations between the linguistic features and readability metrics are investigated. The motivation for this lies in the fact that extraction of the linguistic features is relatively expensive and unreliable, while the computation of the readability metrics is done automatically and with greater accuracy. The ability to estimate the accessibility of documents for people with ASD on the basis of easily computed readability metrics rather than complex linguistic features would be of considerable benefit.

The results obtained in these three parts of the study are presented separately in the following sections.

4.1. Readability

The results of the first part of this study are presented in Table 4. The first row of the table contains the scores for these seven features obtained for SimpleWiki. For the other three text genres, the values of these features were calculated and a non-parametric statistical test (Kolmogorov-Smirnov Z test) was applied in order to calculate the significance of the differences in means between SimpleWiki and the corresponding text genre.[8]

In Table 4, values which differ from those obtained for the documents in SimpleWiki at a 0.01 level of significance are printed in bold. Those printed in bold with an asterisk differ from those obtained from documents in SimpleWiki at a 0.05, but not at a 0.01 level of significance.

On the basis of these results it can be inferred that the news texts are most difficult to read as they require a higher level of literacy for their comprehension (the values of the Kincaid, Fog and SMOG indices are maximal for this genre, while the Flesch index is at its lowest level, indicating that all are in accordance). Discrepancies between the values of different indices are not surprising, as they use different variables and different criterion scores (DuBay, 2004). Also, it is known that the predictions made by these formulae are not perfect, but are rough estimates (r = .50 to .84) of text difficulty. That is, they "account for 50 to 84 percent fo the variance in text difficulty as measured by comprehension tests" (DuBay, 2004). In the context of the current research, it is important that when the difficulty of two types of text is compared, consistent conclusions can be made about which type is more difficult than the other, regardless of which readability formula is used (Table 4).

It is interesting to note that none of the indices indicate significant differences between the readability of health documents and that of documents in SimpleWiki, suggesting that similar levels of literacy are necessary for their comprehension. Despite this, a slightly greater average sentence length was noted for the health texts than the texts from SimpleWiki. The most surprising finding was that the fiction texts are reported by all readability metrics (including average word and sentence length) to be significantly less difficult than those from SimpleWiki (Table 4). These results cast doubt on the assumption that SimpleWiki serves as a paradigm of accessibility, which has been made in previous work on text simplification (e.g. (Coster and Kauchak, 2011)).

As it was observed that all readability indices returned similar values in the comparison of different text genres (Table 4), the strength of correlation between them was investigated. To this end, Pearson's correlation was calculated between each pair of the four indices (Table 6), over the whole corpora (SimpleWiki, News, Health, and Fiction). Pearson's correlation is a bivariate measure of strength of the relationship between two variables, which can vary from 0 (for a random relationship) to 1 (for a perfectly linear relationship) or -1 (for a perfectly negative linear relationship). The results presented in Table 6 indicate a very strong linear correlation between each pair of readability indices.

[8]The Kolmogorov-Smirnov Z test was selected as a result of prior application of the Shapiro-Wilk's W test which demonstrated that most of the features do not follow a normal distribution.

Corpus	V	N	Prep	Det	Adv	Pron	A	CS	CC	INF	Compl	Senses	defNP
SimpleWiki	**2.74**	**5.77**	**1.92**	**1.94**	**0.77**	**0.81**	**1.20**	**0.19**	**0.60**	**0.21**	**1.37**	**6.59**	**1.19**
News	**4.08**	**6.63**	**2.44**	**2.17**	**0.95**	**1.64**	**1.56**	**0.35**	**0.65***	**0.38**	**0.53**	**6.73***	**1.26***
Health	**3.40**	**5.22**	1.82	**1.51**	**0.99**	**1.38**	**1.63**	**0.32**	**1.01**	**0.35**	**1.15**	6.73	**0.73**
Fiction	2.95	**3.33**	**1.43**	**1.35**	**1.10**	**1.89**	**0.90**	**0.23**	**0.49**	**0.23***	**1.13***	**7.59**	**0.77**

Table 5: Linguistic features

r	Kincaid	Flesch	Fog	SMOG
Kincaid	1	−.959	.987	.951
Flesch	−.959	1	−.957	−.972
Fog	.987	−.957	1	.979
SMOG	.951	−.972	.979	1

Table 6: Pearson's correlation between readability indices

In the case of the Flesch index, the higher the score is, the lower the grade level necessary for understanding the given text. For all other indices, a higher score indicates a higher grade level necessary to understand the text. Therefore, the correlation between the Flesch index and any other index is always reported as negative. In order to confirm that these correlations are not genre dependent, a set of experiments was conducted to measure the correlation between these four readability measures separately for each of the four corpora (SimpleWiki, News, Health and Fiction). Those experiments revealed a very close correlation between the four readability indices (between .915 and .993) in each genre.

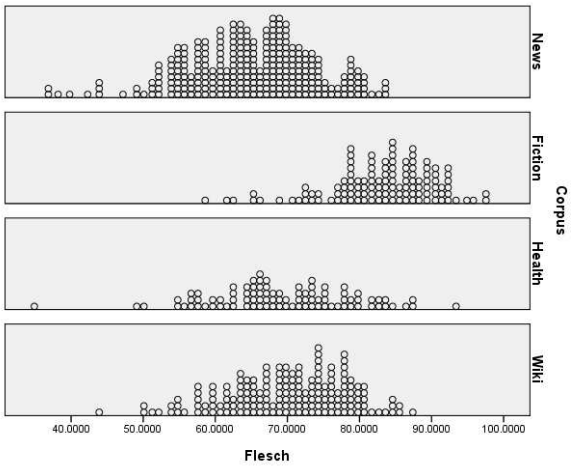


Figure 1: Distribution of the Flesch index

As the correlation between the four readability indices was reported to be almost perfectly linear (Table 6), the reminder of this study focuses on the Flesch index as a representative of the readability indices. The results discussed earlier (Table 4) presented only the mean value of the Flesch index for each of the corpora. In Figure 1, each text is represented separately, providing a more complete picture of the Flesch index distribution across the corpora.

It can be noted that the mean value of the Flesch index lies at approximately the same place on the x-axis for both SimpleWiki and Health texts, which is in accordance with the previously reported results (Table 4). Mean value of the Flesch index in News genre slightly shifted to the left relative to the SimpleWiki corresponds to lower text readability in News genre than in SimpleWiki reported in Table 4. It can also be noted that the distribution of the Flesch index in the Fiction genre is positioned significantly to the right relative to the SimpleWiki, thus indicating a higher readability of texts in this genre.

4.2. Linguistic Features

The investigation of the average occurrence of ten different POS tags per sentence (V[9], N, Prep, Det, Adv, Pron, A, CS, CC, INF) and three other linguistically motivated features (Compl, Senses and defNP) showed significantly different values in News, Health and Fiction than in SimpleWiki in most of the cases (Table 5).

Documents from SimpleWiki were found to contain the *highest* ratio between simple and complex sentences (Compl), the *lowest* number of verbs (V), adverbs (Adv), pronouns (Pron), subordinating conjunctions (CS), infinitive markers (INF) and senses per word (Senses), which may reflect a certain simplicity of these texts.

The News genre was found to contain the *lowest* ratio of simple to complex sentences (Compl), and the *highest* number of verbs (V), subordinate conjunctions (CS) and infinitive markers (INF) per sentence. These features indicate a greater number of verb chains (Compl) and subordinate clauses (CS), longer verb chains and more complex verb constructions (V and INF) for news articles. These features can be considered indicators of syntactic complexity, which is probably reflected in the high scores for readability indices obtained in this genre (Table 4). The texts from the genre of fiction contained the smallest average number of nouns (N), prepositions (Prep), determiners (Det), adjectives (A) and coordinating conjunctions (CC) per sentence (Table 5). However, this genre contained a significantly higher number of senses per word (Senses) than other genres.

4.3. Flesch vs. Linguistic Features

In the third part of the study, potential correlation between the linguistic features and readability indices was investigated. The Flesch index was selected as a representative of readability indices (as all four readability indices were almost perfectly linearly correlated, selection of an alternative readability index should not change the results significantly). Pearson's correlation between the investigated POS frequencies (on average per sentence)

[9]This tag includes the occurrence of present (ING) and past participle (EN).

Corpus	V	N	Prep	Det	Adv	Pron	A	CS	CC	INF
all	**−.493**	**−.812**	**−.777**	**−.715**	**−.093***	**.189**	**−.769**	−.377	**−.464**	**−.415**
SimpleWiki	**−.397**	**−.552**	**−.641**	**−.545**	**−.293**	.136	**−.685**	−.130	**−.424**	−.118
News	**−.385**	**−.738**	**−.759**	**−.705**	**−.197**	**.291**	**−.783**	**−.438**	**−.387**	**−.426**
Health	**−.274**	**−.743**	**−.607**	**−.489**	−.104	.078	**−.703**	.014	**−.610**	−.139
Fiction	**−.605**	**−.889**	**−.854**	**−.851**	**−.555**	−.146	**−.876**	**−.515**	**−.670**	**−.506**

Table 7: Pearson's correlation between Flesch readability index and POS frequencies

Corpus	Compl	ch/w	syl/w	w/s	Senses	defNP
All	**.210**	**−.859**	**−.922**	**−.792**	**.627**	**−.595**
SimpleWiki	**.209**	**−.825**	**−.921**	**−.643**	**.452**	**−.337**
News	−.026	**−.866**	**−.919**	**−.762**	**.568**	**−.688**
Health	0.034	**−.771**	**−.918**	**−.705**	**.417**	**−.450**
Fiction	**.376**	**−.790**	**−.877**	**−.822**	**.738**	**−.791**

Table 8: Pearson's correlation between Flesch readability index and other features

and the Flesch index is presented in Table 7, while the correlation between the other six features and the Flesch index is reported in Table 8. These experiments were conducted first for all the corpora and then for each corpus separately in order to determine whether these correlations may be genre dependent.

As would be expected, the direction of correlation (sign '−' or '+') is independent of genre (in those cases where the correlation is statistically significant and thus more reliable). However, the strength of the correlation does depend of the genre of the texts, e.g. correlation between average number of verbs per sentence (V) and the Flesch index is −.274 for the Health and −.605 for the Fiction genres. The '−' sign indicates that if the value of the feature increases, the Flesch index decreases (indicating a less readable text) and vice-versa (as the Pearson's correlation is a symmetric function we are not able to say in which direction the correlation goes). The results presented in Tables 7 and 8 therefore indicate that for most of the features (V, N, Prep, Det, Adv, A, CS, CC, INF, ch/w, w/s, defNP) the lower the feature value for a given text, the easier that text is to read (the higher the Flesch index). For feature Compl, the results also support the intuition that the higher the ratio of simple to complex sentences is in the text, the more readable it is (higher Flesch index).

The most surprising results were those obtained for the feature Senses (Table 8), which indicate that the higher the average number of senses per word in the text, the more readable the text is. One possible hypothesis that emerges from this observation is that shorter words in English tend to be more semantically ambiguous than longer words (the readability indices are highly correlated with word length, measured both in characters and syllables per word, with the occurrence of shorter words suggesting that the text is easier to read).

5. Conclusions

There are several important findings of this study. First, it was shown that the four well-known readability indices are almost perfectly linearly correlated on each of the four investigated text genres – SimpleWiki, News, Health, and Fiction. Furthermore, our results indicated that texts from the genre of fiction are simpler than those selected from SimpleWiki in terms of the readability indices, casting doubt on the assumption that SimpleWiki is a useful source of documents to form a gold standard of accessibility for people with reading difficulties. Application of the measures also indicated that news articles are most difficult to read, relative to the other genres, requiring a higher level of literacy for their comprehension.

The results of the second part of our study (investigation of various linguistic features) revealed that documents from SimpleWiki were the simplest of the four corpora in terms of several linguistic features – average number of verbs, adverbs, pronouns, subordinate conjunctions, infinitive markers, number of different word senses and ratio between simple and complex sentences. They also indicated some of the factors that may make news texts difficult to read, e.g. containing the highest numbers of verbs and subordinate conjunctions per sentence, and the lowest ratio of simple to complex sentences.

The results of the third set of experiments indicated the average length of words (in characters and in syllables) as being features with the highest correlation to the Flesch index. They also indicated that features such as the average number of nouns, prepositions, determiners and adjectives are closely correlated with the Flesch index (up to .89 in the fiction genre), which supports the idea of using readability indices as an initial measure of text complexity in our project. The comparison of these correlations across different text genres demonstrated that they are genre dependent and that the correlation between these linguistic features and the Flesch index is closest for the Fiction genre.

6. Acknowledgements

The research described in this paper was partially funded by the European Commission under the Seventh (FP7 - 2007-2013) Framework Programme for Research and Technological Development (FIRST 287607). This publication [communication] reflects the views only of the authors, and the Commission cannot be held responsible for

any use which may be made of the information contained therein.

7. References

J. R. Bormuth. 1966. Readability: A new approach. *Reading research quarterly*, 1:79–132.

W. Bosma, 2005. *Image retrieval supports multimedia authoring*, pages 89–94. ITC-irst, Trento, Italy.

L. Burnard. 1995. *Users Reference Guide British National Corpus Version 1.0*. Oxford University Computing Services, UK.

N. Chomsky. 1986. *Knowledge of language: its nature, origin, and use*. Greenwood Publishing Group, Santa Barbara, California.

E. B. Coleman, 1965. *On understanding prose: some determiners of its complexity*. National Science Foundation, Washington, D.C.

E. B. Coleman, 1971. *Developing a technology of written instruction: some determiners of the complexity of prose*. Teachers College Press, Columbia University, New York.

W. Coster and D. Kauchak. 2011. Simple english wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, pages 665–669, Portland, Oregon, June. Association of Computational Linguistics.

C. C. Doak, L. G. Doak, and J. H. Root. 1996. *Teaching patients with low literacy skills*. J. B. Lippincott Company, Philadelphia.

W. H. DuBay. 2004. *The Principles of Readability*. Impact Information, Costa Mesa.

G. Escudero, L. Màrquez, and G. Rigau. 2000. A comparison between supervised learning algorithms for word sense disambiguation. In C. Cardie, W. Daelemans, C. Nédellec, and E. Tjong Kim Sang, editors, *Proceedings of the Fourth Computational Natural Language Learning Workship, CoNLL-2000*, pages 31–36, Lisbon, Portugal, September. Association of Computational Linguistics.

R. Evans. 2011. Comparing methods for the syntactic simplification of sentences in information extraction. *Literary and Linguistic Computing*, 26 (4):371–388.

C. Fellbaum. 1998. *WordNet An Electronic Lexical Database*. MIT Press, Cambridge, MA.

R. Flesch. 1949. *The art of readable writing*. Harper, New York.

U. Frith and M. Snowling. 1983. Reading for meaning and reading for sound in autistic and dyslexic children. *Journal of Developmental Psychology*, 1:329–342.

R. Gaizauskas, J. Foster, Y. Wilks, J. Arundel, P. Clough, and S. Piao. 2001. The Meter corpus: A corpus for analysing journalistic text reuse. In *Proceedings of Corpus Linguistics 2001 Conference*, pages 214–223. Lancaster University Centre for Computer Corpus Research on Language.

W. S. Gray and B. Leary. 1935. *What makes a book readable*. Chicago University Press, Chicago.

R. Gunning. 1952. *The technique of clear writing*. McGraw-Hill, New York.

R. P. Kern. 2004. *Usefulness of readability formulas for achieving Army readability objectives: Research and state-of-the-art applied to the Army's problem (NTIS No. AD A086 408/2)*. U.S. Army Research Institute, Fort Benjamin Harrison.

J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1986. *Derivation of new readability formulas (Automatic Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. CNTECHTRA.

G. H. McLaughlin. 1969. SMOG grading – a new readability formula. *Journal of reading*, 22:639–646.

N. Minshew and G. Goldstein. 1998. Autism as a disorder of complex information processing. *Mental Retardation and Developmental Disability Research Review*, 4:129–136.

R. Mitkov. 2002. *Anaphora Resolution*. Longman, Harlow, Essex.

K. Nation, P. Clarke, B. Wright, and C. Williams. 2006. Patterns of reading ability in children with autism-spectrum disorder. *Journal of Autism & Developmental Disorders*, 36:911–919.

NCES. 1993. *Adult literacy in America*. National Center for Education Statistics, U.S. Dept. of Education, Washington, D.C.

I. M. O'Connor and P. D. Klein. 2004. Exploration of strategies for facilitating the reading comprehension of high-functioning students with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 34:2:115–127.

C. Orăsan and L. Hasler. 2007. Computer-aided summarisation: how much does it really help? In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, pages 437–444, Borovets, Bulgaria, September.

A. Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:1:77–109.

P. Tapanainen and T. Jarvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th conference on Applied Natural Language Processing of the Association for Computational Linguistics*, pages 64–71. Association of Computational Linguistics.

B. D. Weiss and C. Coyne. 1997. The use of user modelling to guide inference and learning. *New England Journal of Medicine*, 337 (4):272–274.

M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from wikipedia. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pages 365–368, Los Angeles, California, June. Association of Computational Linguistics.

A First Approach to the Creation of a Spanish Corpus of Dyslexic Texts

Luz Rello[1,2], Ricardo Baeza-Yates[3,2], Horacio Saggion[1], Jennifer Pedler[4]

[1]NLP and [2]Web Research Groups, Universitat Pompeu Fabra, Barcelona, Spain
[3]Yahoo! Research, Barcelona, Spain
[4]Dept. of Computer Science, Birkbeck, University of London
luzrello@acm.org, rbaeza@acm.org, horacio.saggion@upf.edu, jenny@dcs.bbk.ac.uk

Abstract

Corpora of dyslexic texts are valuable for studying dyslexia and addressing accessibility practices, among others. However, due to the difficulty of finding texts written by dyslexics, these kind of resources are scarce. In this paper, we introduce a small Spanish corpus of dyslexic texts with annotated errors. Since these errors require non-standard annotation, we present the annotation criteria established for the different types of dyslexic errors. We compare our preliminary findings with a similar corpus in English. This comparison suggests that the corpus shall be enlarged in future work.

Keywords: Corpus, Non-standard annotation, Errors, Dyslexia.

1. Introduction

Worldwide, around 15-20% of the population has a language-based learning disability; where 70-80% of them are likely dyslexic (International Dyslexia Association, 2011).

Regarding this substantial group of people, various accessibility studies take dyslexia into account. They mainly focus on tools (Pedler, 2007; Gregor et al., 2003) and guidelines for dyslexic-accessible practices (McCarthy and Swierenga, 2010). There is a common agreement in these studies that the application of dyslexic-accessible practices benefits also the readability for non-dyslexic users as well as other users with disabilities such as low vision (Evett and Brown, 2005).

Although the use of corpora of dyslexic errors have been used for various purposes such as diagnosing dyslexia (Schulte-Körne et al., 1996) and developing tools, i.e. spell checkers (Pedler, 2007), their existence is scarce.

In this paper we present the following contributions:

- The first approach to create a corpus of dyslexic errors in Spanish,
- guidelines for the annotation of dyslexic errors and,
- a comparison of our corpus with a similar corpus in English.

In the next section we make a brief explanation of dyslexia and explain in Section 3 how dyslexic errors have been used for different purposes. In Section 4 we describe our related work, Pedler's corpus of dyslexic texts in English (Pedler, 2007), and in Section 5 we present a classification of the dyslexic errors. Sections 6 and 7 detail the characteristics of our corpus and its annotation guidelines. In Section 8 we compare the distribution of dyslexic errors in English and Spanish. Conclusions and future work are drawn in Section 9.

2. What is Dyslexia?

Dyslexia is a specific learning disability which is neurological in origin. It is characterized by difficulties with accurate and/or fluent word recognition and by poor spelling and decoding abilities. These difficulties typically result from a deficit in the phonological component of language that is often unexpected in relation to other cognitive abilities. Secondary consequences may include problems in reading comprehension and reduced reading experience that can impede growth of vocabulary and background knowledge (Lyon et al., 2003; Orton Dyslexia Society Research Committee, 1994).

In some literature, dyslexia is referred to as a specific reading disability (Vellutino et al., 2004) and dysgraphia its writing manifestation (Romani et al., 1999).[1] However, our study follows the standard definitions of ICD-10 and DSM-IV (World Health Organization, 1993; American Psychiatric Association, 2000) where dyslexia is listed as a reading and spelling disorder.

Despite its universal neurocognitive basis, dyslexia manifestations are variable and culture-specific (Goulandris, 2003). This variability is due to the different language orthographies concerning their grade of consistency and regularity (Brunswick, 2010). English has an opaque –or deep– orthography in which the relationships between letters and sounds are inconsistent and many exceptions are permitted. English presents a significantly greater challenge to the beginning reader than other languages, such as Spanish, with a more regular alphabetic system that contains consistent mappings between letters and sounds, that is, a transparent –or shallow– orthography.

Depending on the language, the estimations on the prevalence of dyslexia varies. The (Interagency Commission on Learning Disabilities, 1987) states that 10-17.5% of the population in the U.S.A. has dyslexia. The model of Shaywitz *et al.* (1992) predicts that 10.8% of English speaking children have dyslexia while in (Katusic et al., 2001) the rates varied from 5.3% to 11.8% depending on the formula used.

[1]Dysgraphia refers to a writing disorder associated with the motor skills involved in writing, handwriting and sequencing, but also orthographic coding (Romani et al., 1999). It is comorbid with dyslexia, that is, it is a medical condition that co-occurs with dyslexia (Nicolson and Fawcett, 2011).

3. The Use of Dyslexic Errors

In general terms, errors could be used as a source of knowledge. For instance, the presence of errors in the textual Web have been used for detecting spam (Piskorski et al., 2008), measuring quality (Gelman and Barletta, 2008) and understandability (Rello and Baeza-Yates, 2012) of web content. Among the different kind of errors found in the Web, at least 0.67% errors are only made by dyslexic users (Baeza-Yates and Rello, 2011). In the case of people with dyslexia, their written errors have been used for various accessibility related purposes such as the development of tools like spell checkers (Pedler, 2007) or word processors (Gregor et al., 2003).

Besides the accessibility practices, analyses of writing errors made by dyslexics have been used in previous literature to study different aspects of dyslexia. For instance, the specific types of dyslexic errors highlight different aspects of dyslexia (Treiman, 1997) such as a phonological processing deficit (Moats, 1996; Lindgrén and Laine, 2011). People with dyslexia exhibit higher spelling error rates than non-dyslexic people (Coleman et al., 2009) and, due to this fact, there are diagnosis of dyslexia based on the spelling score (Schulte-Körne et al., 1996). According to (Meng et al., 2005) only 30% of dyslexics have trouble with reversing letters and numbers. However, errors attributable to phonological impairment, spelling knowledge, and lexical mistakes are more frequent in dyslexics than in non-dyslexics (Sterling et al., 1998). Nonetheless, the dyslexic error rate vary depending on the language writing system (Lindgrén and Laine, 2011).

4. Related Work

To the best of our knowledge, there is only one corpus of dyslexic texts, the corpus used by Pedler (2007) for the creation of a spell checker of real-word errors made by dyslexic people.

This corpus in English is composed of 3,134 words and 363 errors (Pedler, 2007). This corpus is made of: (1) word-processed homework (saved before it was spellchecked) produced by a third year secondary school student; (2) two error samples used for a comparative test of spellcheckers (Mitton, 1996); and (3) short passages of creative writing produced by secondary school children of low academic ability in the 1960s (Holbrook, 1964).

To develop a program designed to correct actual errors made by dyslexics, this initial corpus was enlarged to 12,000 words containing just over 800 real-word errors.[2] The additional sources for that corpus were: texts from a dyslexic student, texts from an online typing experiment (Spooner, 1998), samples from dyslexic bulletin boards and mailing lists and stories written by dyslexic children.

All the errors in this corpus were annotated in the format illustrated next, where **pituwer* is the dyslexic error from the intended work *picture*.[3]

```
<ERR targ=picture> pituwer </ERR>
```

Our current annotation method is inspired by Pedler's work (2007) and is described in Section 7.

5. Types of Dyslexic Errors

Pedler (2007) found the following kinds of dyslexic errors in her corpus and proposed the following classification of dyslexic errors:

1. Dyslexic errors based on the degree of difference to the intended or target word:
   (a) Simple errors. They differ from the intended word by only a single letter. They can be due to:
      i. substitution, **reelly (really)*,
      ii. insertion, **situartion (situation)*,
      iii. omission, **approch (approach)* and
      iv. transposition, **artcile (article)*.

      In (Damerau, 1964), 80% of the misspellings in his corpus (non-dyslexic errors) were simple errors.[4]
   (b) Multi-errors. They differ in more than one letter from the target word. Some errors, such as **queraba (quedara, 'stayed')*, closely resemble the intended word, while others are not so obvious, **lignsuitc (linguistics)*.
   (c) Word boundary errors. They are mistakes (run-ons and split words) which are special cases of omission and insertion errors. A run-on is the result of omitting a space, such as **alot (a lot)* while a split word occurs when a space is inserted in the middle of a word, such as **sub marine (submarine)*.
2. Dyslexic errors based on their correspondence with existing words:
   (a) Real-word errors. Misspellings that result in another valid word. For instance, *witch* being the intended word *which*.
   (b) Non-word errors. Misspellings that do not result in another correct word, such as **conmitigo (contigo, 'with you')*
3. First letter dyslexic errors:
   (a) First letter errors, like **no (know)*.

6. Spanish Corpus of Dyslexic Texts

Manifestations of dyslexia varies among languages (Goulandris, 2003) but also among subjects and among ages (Vellutino et al., 2004). For instance misspelling rate in dyslexic children is higher than in adults (Sterling et al.,

[2] A corpus containing 833 dyslexic real-word errors in context is available at: `http://www.dcs.bbk.ac.uk/~jenny/resources.html`

[3] Dyslexic errors are preceded by * while the intended target word follows in parenthesis.

[4] The standard definition of edit distance (Levenshtein, 1965) consider transpositions as two errors, while Damerau defined it as a single error.

1998). However, experiments evidence that adult dyslexics have a continuing problem in the lexical domain, manifested in poor spelling ability (Sterling et al., 1998).
Due to this variability, we pursued to collect texts written by a similar population in terms of age, education, native language and diagnosed dyslexia. We collected 16 Spanish texts written by dyslexic children from 13 to 15 years old. The texts are composed of homework writing exercises and were written by children who had Spanish as native language. The texts were all handwritten and we transcribed them manually. The words that we were not able to transcript due to the illegibility of the hand writing were marked. One example of a fragment of our texts is given in Figure 1.

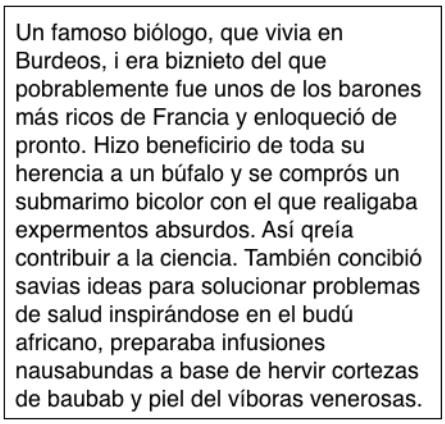

Un famoso biólogo, que vivia en Burdeos, i era biznieto del que pobrablemente fue unos de los barones más ricos de Francia y enloqueció de pronto. Hizo beneficirio de toda su herencia a un búfalo y se comprós un submarimo bicolor con el que realigaba expermentos absurdos. Así qreía contribuir a la ciencia. También concibió savias ideas para solucionar problemas de salud inspirándose en el budú africano, preparaba infusiones nausabundas a base de hervir cortezas de baubab y piel del víboras venerosas.

Figure 1: Example of one story of the texts written by a dyslexic child (14 years).

In the example in Figure 1[5] we have errors of all possible kinds, most of them simple: (i) substitution: **i (y), *realigaba (realizaba), *qreía (creía), *savias (sabias), *budú (vudú), *venerosas (venenosas)* and **baubab (baobab).*; (ii) insertion: **comprós (compró),* ; (iii) omission: **expermentos (experimentos), *unos (uno), *beneficirio (beneficiario), *nausabundas (nauseabundas)* and **del (de)*; and a double (iv) transposition **pobrablemente (probablemente)*.
We observe that there are errors that might not be attributed to dyslexia, for instance **i (y)* could be easily attributed as a transference from Catalan language (bilingual writer) and two others are concordance errors (**unos* and **del*). There is also one accentuation error: **vivia (vivía)*.
Since dyslexic errors overlap with other kind of errors found in documents, it is challenging to determine which errors are more likely to be only done by dyslexics. However, non-word multi-errors are more likely to be produced by a person with dyslexia (Baeza-Yates and Rello, 2011).

[5]Approximated literal translation: A famous biologist, who lived in Bordeaux, and was great-grandson of who probably was one of the wealthiest barons of France and suddenly went mad. He chose a buffalo as the beneficiary of his inheritance and bought a bicolor submarine in which he made absurd experiments. So he thought that he contributed to science. He also conceived wise ideas to solve health problems inspired by the African voodoo, preparing nauseating infusions based on boiled baobab barks and poisonous snakes.

The length average per text is 67 words and the total corpus size is 1,057 words. The reduced size of the corpus is explained by the difficulty of finding texts written by people diagnosed with dyslexia and the lack of a previous Spanish corpus of dyslexic errors. However, we believe that a corpus of this characteristics is valuable to analyze Spanish dyslexic errors and provide insight in where they appear or which is their distribution in Spanish. To the best of our knowledge, lists but not texts of dyslexic errors were used in previous work (Silva Rodríguez and Aragón Borja, 2000; Baeza-Yates and Rello, 2011).

7. Annotation of Dyslexic Errors

Following Pedler's annotation tag for errors, we marked-up all the errors in XML format. This kind of simple annotation gives the possibility, using regular expressions, to extract the errors and their corresponding target word from the corpus, as well as computing statistics.
We manually annotated the errors and added several tag attributes to typify each dyslexic error. Following we present the attributes and their possible values.

- Targ: the correct word(s).
- Type: this attribute refers to the error type depending on their edit distance. Its possible values are: "simple", "multi" and "boundary". Boundary specifies the case when one word is slit or two words are joined.
- Real: this attribute records if the error produced another real word. These errors are the most difficult to find automatically.
- First Letter: if the error is in the first letter or not.
- Edit Distance: The edit distance to the correct word(s).

Below we show an example for the error **pobrablemente (probablemente)* ('maybe').

```
<ERR targ = "probablemente"
type = "multi"
real = "no"
first_letter = "no"
ed = "2" >
pobrablemente </ERR>
```

In the case that there were two kind of errors we annotated as a multi-error, for instance, in **devidreo (de vidrio)* ('of glass') a boundary error is combined with a simple substitution error.
We did not annotate capitalization errors and accentuation errors since children among that age are still learning how to accentuate in Spanish. If the handwriting word was illegible an empty tag `<ILLEGIBLE/>` was added.

8. Comparing English and Spanish Errors

The corpora that we compare in this paper are in English and Spanish. These languages are archetypes of deep and shallow orthographies, respectively. Along an orthographic transparency scale for European languages, English appears as the language with the deepest orthography and

Spanish as the second most shallow after Finnish (Seymour et al., 2003).
In Tables 1 and 2 we compare the data of the corpus described in (Pedler, 2007) with our corpus. We compute the error ratio as the fraction of errors over the correctly spelt words we observe. As expected, Spanish dyslexics make less spelling errors (15%) than English dyslexics (20%) due to their different orthographies. On the other hand the percentage of unique errors is almost the same.

Category	English	Spanish
Total words	3,134	1,075
Total errors	636	157
Error ratio	0.20	0.15
Distinct errors	577	144
Percentage	90.7	91.7

Table 1: Error ratio and percentage in English and Spanish corpora of dyslexic errors.

Table 2 presents the distribution the different types of dyslexic errors for both corpus. To determine if an error was a real world error we checked its existence in the Royal Spanish Academy Dictionary (Real Academia Española, 2001).

Category	English		Spanish	
Simple errors	307	53%	96	67%
Multi errors	227	39%	33	23%
Word boundary errors	47	8%	15	10%
Real-word errors	100	17%	30	21%
Non-word errors	477	83%	114	79%
First letter errors	30	5%	16	11%
Total	577	100%	144	100%

Table 2: Error distribution in English an Spanish corpora of dyslexic errors.

As expected, there is a greater percentage of multi errors in a language with deep orthography as English than in Spanish, i.e. **qreía (creía)* ('thought'). However, the first letter errors are double in Spanish, i.e.: **tula (ruta)* ('way'). This is surprising according to (Yannakoudakis and Fawthrop, 1983) whose findings report that the first letter of a misspelling is correct in the majority of cases.
The rest of the dyslexic error types are similar in both languages. There are slightly more real word errors in Spanish, **dijo (digo)* ('said') or **llegada (llegaba)* ('said').
Simple errors are the most frequent ones in both languages. However, each error type has a different frequency. For instance, in our corpus substitution errors, **dertro (dentro)* ('in') are the most frequent ones (65% of the simple errors) while (Bustamante and Díaz, 2006) states that simple omissions are the most frequent kind.

9. Conclusions and Future Work

The comparisons presented in this works among different kind of dyslexic errors shed light on how dyslexia manifestations varies among languages and suggest that dyslexic accessible practices and tools are partially language dependent. This corpus is available for the research community.[6] Due to the difficulty of collecting texts of diagnosed dyslexics our Spanish corpus is still small but enough to present the distribution of the dyslexic errors and to settle the annotation criteria. In future work we plan to enlarge this corpus with more texts written by dyslexics and also using the Web as corpus. Also we plan to improve its annotation by separating the number of errors (simple or multi) from the case of happening at the boundaries of a word as simple and multi errors overlap with word boundary errors.

Acknowledgements

We deeply thank Yolanda Otal de la Torre for helping us to collect the Spanish texts written by dyslexics.

10. References

American Psychiatric Association. 2000. *Diagnostic and statistical manual of mental disorders: DSM-IV-TR*. American Psychiatric Publishing, Inc.

R. Baeza-Yates and L. Rello. 2011. Estimating dyslexia in the Web. In *International Cross Disciplinary Conference on Web Accessibility (W4A 2011)*, pages 1–4, Hyderabad, India, March. ACM Press.

Nicola Brunswick. 2010. Unimpaired reading development and dyslexia across different languages. In Sine McDougall and Paul de Mornay Davies, editors, *Reading and dyslexia in different orthographies*, pages 131–154. Psychology Press, Hove.

F. R. Bustamante and E.L. Díaz. 2006. Spelling error patterns in spanish for word processing applications. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 93–98. ELRA.

C. Coleman, N. Gregg, L. McLain, and L. W. Bellair. 2009. A comparison of spelling performance across young adults with and without dyslexia. *Assessment for Effective Intervention*, 34(2):94–105.

F.J. Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the A.C.M.*, 7:171–176.

L. Evett and D. Brown. 2005. Text formats and web design for visually impaired and dyslexic readers-clear text for all. *Interacting with Computers*, 17:453–472, July.

I. A. Gelman and A. L. Barletta. 2008. A "quick and dirty" website data quality indicator. In *The 2nd ACM workshop on Information credibility on the Web (WICOW '08)*, pages 43–46.

N.E. Goulandris. 2003. *Dyslexia in different languages: Cross-linguistic comparisons*. Whurr Publishers.

P. Gregor, A. Dickinson, A. Macaffer, and P. Andreasen. 2003. Seeworda personal word processing environment for dyslexic computer users. *British Journal of Educational Technology*, 34(3):341–355.

D. Holbrook. 1964. English for the rejected: Training literacy in the lower streams of the secondary school.

Interagency Commission on Learning Disabilities. 1987. *Learning Disabilities: A Report to the U.S. Congress*. Government Printing Office, Washington DC, U.S.

[6] http://www.luzrello.com/Dyswebxia.html

International Dyslexia Association. 2011. Frequently Asked Questions About Dyslexia. http://www.interdys.org/.

S.K. Katusic, R.C. Colligan, W.J. Barbaresi, D.J. Schaid, and S.J. Jacobsen. 2001. Incidence of reading disability in a population-based birth cohort, 1976-1982, rochester, mn. *Mayo Clinic Proceedings*, 76(11):1081.

V. Levenshtein. 1965. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17.

S.A. Lindgrén and M. Laine. 2011. Multilingual dyslexia in university students: Reading and writing patterns in three languages. *Clinical Linguistics & Phonetics*, 25(9):753–766.

G.R. Lyon, S.E. Shaywitz, and B.A. Shaywitz. 2003. A definition of dyslexia. *Annals of Dyslexia*, 53(1):1–14.

Jacob E. McCarthy and Sarah J. Swierenga. 2010. What we know about dyslexia and web accessibility: a research review. *Universal Access in the Information Society*, 9:147–152, June.

H. Meng, S. Smith, K. Hager, M. Held, J. Liu, R. Olson, B. Pennington, J. DeFries, J. Gelernter, T. O'Reilly-Pol, S. Somlo, P. Skudlarski, S. Shaywitz, B. Shaywitz, K. Marchione, Y. Wang, P. Murugan, J. LoTurco, P. Grier, and J. Gruen. 2005. DCDC2 is associated with reading disability and modulates neuronal development in the brain. *Proceedings of the National Academy of Sciences*, 102:17053–17058, November.

R. Mitton. 1996. *English spelling and the computer*. Longman Group.

L.C. Moats. 1996. Phonological spelling errors in the writing of dyslexic adolescents. *Reading and Writing*, 8(1):105–119.

R.I. Nicolson and A.J. Fawcett. 2011. Dyslexia, dysgraphia, procedural learning and the cerebellum. *Cortex*, 47(1):117–127.

Orton Dyslexia Society Research Committee. 1994. Definition of dyslexia. Former name of the International Dyslexia Association.

J. Pedler. 2007. *Computer Correction of Real-word Spelling Errors in Dyslexic Text*. Ph.D. thesis, Birkbeck College, London University.

Jakub Piskorski, Marcin Sydow, and Dawid Weiss. 2008. Exploring linguistic features for web spam detection: a preliminary study. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, AIRWeb '08, pages 25–28, New York, NY, USA. ACM.

Real Academia Española. 2001. *Diccionario de la lengua española*. Espasa-Calpe, Madrid, 22 edition.

L. Rello and R. Baeza-Yates. 2012. Lexical quality as a proxy for web text understandability. In *The 21st International World Wide Web Conference (WWW 2012)*, April.

C. Romani, J. Ward, and A. Olson. 1999. Developmental surface dysgraphia: What is the underlying cognitive impairment? *The Quarterly Journal of Experimental Psychology*, 52(1):97–128.

G. Schulte-Körne, W. Deimel, K. Müller, C. Gutenbrunner, and H. Remschmidt. 1996. Familial aggregation of spelling disability. *Journal of Child Psychology and Psychiatry*, 37(7):817–822.

P.H.K. Seymour, M. Aro, and J.M. Erskine. 2003. Foundation literacy acquisition in european orthographies. *British Journal of psychology*, 94(2):143–174.

A. Silva Rodríguez and L.E. Aragón Borja. 2000. Análisis cualitativo de un instrumento para detectar errores de tipo disléxico (IDETID-LEA). *Psicothema*, 12(2):35–38.

R. Spooner. 1998. *A spelling aid for dyslexic writers*. Ph.D. thesis, PhD thesis, University of York.

C. Sterling, M. Farmer, B. Riddick, S. Morgan, and C. Matthews. 1998. Adult dyslexic writing. *Dyslexia*, 4(1):1–15.

R. Treiman. 1997. Spelling in normal children and dyslexics. *Foundations of reading acquisition and dyslexia: Implications for early intervention*, pages 191–218.

F.R. Vellutino, J.M. Fletcher, M.J. Snowling, and D.M. Scanlon. 2004. Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of child psychology and psychiatry*, 45(1):2–40.

World Health Organization. 1993. *International statistical classification of diseases, injuries and causes of death (ICD-10)*. World Health Organization, tenth edition.

E.J. Yannakoudakis and D. Fawthrop. 1983. The rules of spelling errors. *Information Processing & Management*, 19(2):87–99.

Towards Shallow Grammar Induction for an Adaptive Assistive Vocal Interface: a Concept Tagging Approach

Janneke van de Loo[1], Guy De Pauw[1], Jort F. Gemmeke[2], Peter Karsmakers[3], Bert Van Den Broeck[3], Walter Daelemans[1], Hugo Van hamme[2]

[1]CLiPS - Computational Linguistics
University of Antwerp
Antwerp, Belgium
janneke.vandeloo@ua.ac.be
guy.depauw@ua.ac.be
walter.daelemans@ua.ac.be

[2]ESAT - PSI Speech Group
KU Leuven
Leuven, Belgium
jort.gemmeke@esat.kuleuven.be
hugo.vanhamme@esat.kuleuven.be

[3]MOBILAB
K.H. Kempen
Geel, Belgium
peter.karsmakers@khk.be
bert.van.den.broeck@khk.be

Abstract

This paper describes research within the ALADIN project, which aims to develop an adaptive, assistive vocal interface for people with a physical impairment. One of the components in this interface is a self-learning grammar module, which maps a user's utterance to its intended meaning. This paper describes a case study of the learnability of this task on the basis of a corpus of commands for the card game *patience*. The collection, transcription and annotation of this corpus is outlined in this paper, followed by results of preliminary experiments using a shallow concept-tagging approach. Encouraging results are observed during learning curve experiments, that gauge the minimal amount of training data needed to trigger accurate concept tagging of previously unseen utterances.

1. Introduction

Voice control of devices we use in our daily lives is still science fiction: we do not talk to elevators, fridges or heaters. The main reason for this poor market penetration is that often more straightforward alternatives are available, such as pushing a button or using remote controls. Furthermore, speech recognition still lacks robustness to speaking style, regional accents and noise, so that users are typically forced to adhere to a restrictive grammar and vocabulary in order to successfully *command and control* a device. In a commercial climate that focuses on the development of plug 'n play, user-friendly devices, users are loath to adapt to their equipment by reading manuals or documentation or by following training.

But what if pushing buttons is not trivial? Physically impaired people with restricted (upper) limb motor control are permanently in the situation where voice control could significantly simplify some of the tasks they want to perform (Noyes and Frankish, 1992). By regaining the ability to control more devices in the living environment, voice control contributes to their independence of living, their security, their quality of life, their communicative abilities and their entertainment.

The ALADIN project[1] aims to develop a command and control interface for people with a physical impairment, using technology based on *learning* and *adaptation*: the interface should **learn** what the user means with commands, which words he/she uses and what his/her vocal characteristics are. Users should formulate commands as they like, using the words and grammatical constructs they like and only addressing the functionality they are interested in.

The language independent ALADIN system will contain two modules that reduce the amount of linguistic adaptation required from the user:

- The **word finding** module works on the acoustic level and attempts to automatically induce the vocabulary of the user during training, by associating acoustic patterns (command) with observed changes in the user's environment (control).
- The **grammar induction** module works alongside the word finding module to automatically detect the compositionality of the user's utterances, further enabling the user to freely express commands in their own words.

This paper describes work on a self-learning grammar module for the ALADIN interface. A grammar module for a command & control interface enables a mapping between the structural, grammatical properties of a user's utterance and the semantic content of the utterance, i.e. the intended control. Traditionally, command & control interfaces may include a context-free grammar, as illustrated in Figure 1, for the task of operating a television set. The compositionality of possible commands are strictly defined in this grammar, as well as their association with the intended controls (indicated between square brackets).

The ALADIN grammar module, however, will attempt to automatically derive the compositionality of the commands, while keeping the training phase as brief as possible. In this paper, we will outline preparatory experiments towards achieving this goal: before attempting *unsupervised* (shallow) grammar induction of ASR output, this paper will first investigate the feasibility of the induction task itself. The grammar module is investigated in isolation and under ideal circumstances, i.e. using manually transcribed and annotated data. Section 2 will describe the task at hand and the annotated corpus developed to investigate the aforemen-

[1]Adaptation and Learning for Assistive Domestic Vocal INterfaces. Project page:
http://www.esat.kuleuven.be/psi/spraak/projects/ALADIN

```
<sentence>         = <volume_command> | <channel_command>
<volume_command>   = (set | change) volume [VOL] to <number>
<channel_command>  = (select | change to) channel [CH] (<number> | <name>)
<number>           = one [1] | two [2] | three [3] | four [4] | five [5]
<name>             = BBC [4] | CNN [2] | EuroSports [1]
```

Figure 1: Context-free grammar for a television command & control interface.

tioned research goals. Section 3 outlines the envisioned approach, i.e. concept tagging. The paper concludes with a discussion of the results and pointers towards future research.

2. Patience Corpus

For many command & control (henceforth C&C) domotica tasks, a grammar is not a strictly necessary commodity. It is perfectly feasible to control your television set using holistic commands for which no compositionality as defined in a grammar (cf. Figure 1) is needed. Furthermore, in case of a command such as "*turn the TV a bit louder*", even the unordered collection of the keywords in the utterance and their associated meanings is usually sufficient to *understand* the utterance and trigger the intended control.

There are however plenty of C&C applications for which knowledge of the compositionality of the utterance is needed to determine its meaning, such as voice controlled GPS systems, controlling entertainment centers and various types of gaming applications. To study the expedience, as well as the feasibility of grammar induction for a manageable, yet non-trivial C&C task, we decided on a case study for the card game of "patience".

Patience (also know as "solitaire") is one of the most well-known single player card games. The playing field (cf. Figure 2) consists of seven columns, four *foundation stacks* (top) and the remainder of the deck, called the *hand* (bottom). The object of the game is to move all the cards from the hand and the seven columns to the foundation stack, through a series of manipulations, in which consecutive cards of alternating colors can be stacked on the columns and consecutive cards of the same suit are placed on the foundation stack.

This game presents an interesting case study, since a C&C interface for this game needs to learn a non-trivial, but fairly restrictive vocabulary and grammar. Commands such as "*put the four of clubs on the five of hearts*" or "*put the three of hearts in column four*" are not replaceable by holistic commands and identifying the individual components of the utterance and their interrelation, is essential for the derivation of its meaning. In this section, we will describe the collection and annotation of a modestly sized corpus of spoken commands for the card game of patience.

2.1. Data Collection

The patience corpus consists of more than two thousand spoken commands in (Belgian) Dutch[2], transcribed and annotated with *concept tags* (cf. Section 3). During data collection, eight participants were asked to play patience on a computer using spoken commands. These commands were subsequently executed by the experimenter. The participants were told to advance the game by using their own commands freely, both in terms of vocabulary and grammatical constructs. The audio signals of the commands were recorded and the associated actions, executed by the experimenter, were stored in the form of action frames (cf. Section 2.2).

During the patience games, the experimenter executing the commands was situated in a separate room, invisible to the participant, and the participant gave commands through a headset microphone. Half of the participants were given the impression that their commands were executed by a completely automated system (Wizard-of-Oz), while the other four participants were told in advance that the commands would be executed by a person.

Both setups approximate the ALADIN C&C situation in their own way: the Wizard-of-Oz setup accounts for effects of human-machine interaction, but inclines people to adapt their commands to what they think the computer program would *understand*, whereas the non-Wizard-of-Oz setup gives people more sense of freedom in making up their own commands, but might also yield commands that people would not use when talking to a computer. We decided to use both setups, in order to obtain a wide range of possible commands. A preliminary qualitative inspection of the corpus did not uncover significant grammatical differences between the two groups of participants, however.

Each participant played in two separate sessions, with at least three weeks in between, so as to capture potential variation in command use over time. The participants' ages range between 22 and 73 and we balanced for gender and education level. We collected between 223 and 278 commands (in four to six games) per participant. The total number of collected commands is 2020, which means an average of 253 commands per participant and the average number of moves per game is 55.

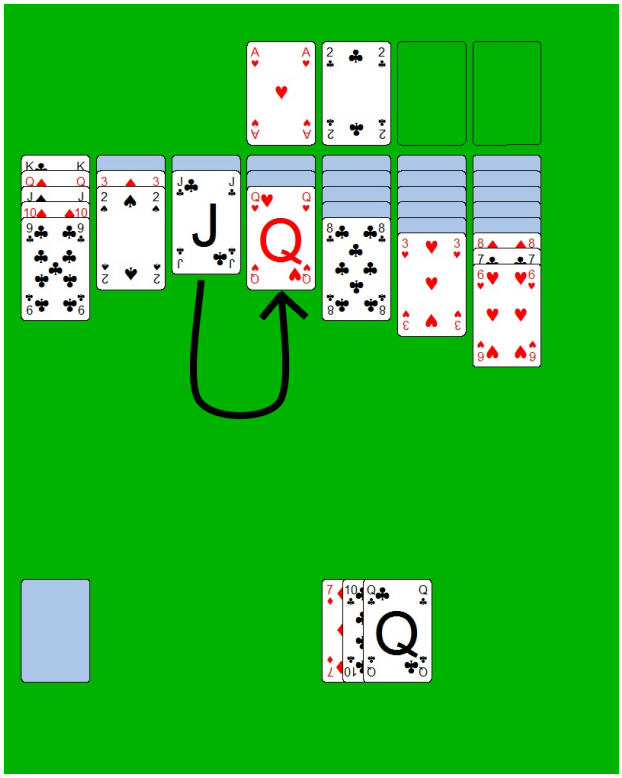
2.2. Action Frames

Each action in the C&C patience implementation was automatically stored in the form of an *action frame*. An action frame is a data structure that represents the semantic concepts that are relevant to the execution of the action and which users of the C&C application are likely to refer to in their commands. Such frame-based semantic representations have previously been successfully deployed in C&C applications and spoken dialog systems (Wang and Acero, 2005; Wang and Acero, 2006).

A frame usually contains one or multiple slots, associated with values. The slots in an action frame represent relevant properties of the action. The patience game has two

[2]Note however that the ALADIN system is inherently language independent.

Leg de klaveren boer op de harten vrouw
(Put the jack of clubs on the queen of hearts)



Frame Slot	Value
<from_suit>	c
<from_value>	11
<from_foundation>	-
<from_column>	3
<from_hand>	-
<to_suit>	h
<to_value>	12
<to_foundation>	-
<to_foundationempty>	-
<to_column>	4
<to_columnempty>	-

Figure 2: An example of a command, the associated action on the screen and the automatically generated `movecard` action frame.

types of action frames: `dealcard` and `movecard`. The former frame type does not have any slots: merely selecting this frame is sufficient for the execution of the action. The `movecard` frame on the other hand does have slots, specifying which card should be moved and to which position it should be moved. Figure 2 shows an example of a command, the action performed on the playing field and the frame description of that action.

Each card is defined as the combination of a `suit`[3] and a `value`[4]. The positions of the cards on the playing field are also represented by the frame description and different stacks are discerned: the hand, at the bottom, containing

[3]hearts(h), diamonds(d), clubs(c) or spades(s).

[4]From ace (1) to king (13).

Frame Slot	Value
<from_suit>	c
<from_value>	11
<to_suit>	h
<to_value>	12

Figure 3: Oracle Command Frame (`movecard`) for the utterance "*put the jack of clubs on the queen of hearts*".

the visible cards which have not been played yet, the seven columns in the center of the playing field, and the four foundation stacks at the top right, where all cards should finally be moved to, ordered by suit.

The `movecard` frame has `from` slots, identifying the card (`suit` and `value`) that is moved and the position (the hand, a column or a foundation stack) from which it is moved, and `to` slots, identifying the card and position that it is moved to. If the card is moved to an empty column, the slot `to_columnempty` is filled with the value `1`. Similarly, the slot `to_foundationempty` receives the value `1` when a card is moved to an empty foundation stack.

Note that the frame description in Figure 2 is over-specified with respect to the actual command. While the command may for instance only mention the cards involved in the move, column numbers are also specified in the frame description. This is due to the fact that the program generates the frame descriptions without any knowledge of the actual audio or its content. The final grammar module will therefore need to be able to not only identify the compositionality of the utterance, but also which subset of frame slots are actually mentioned by the user.

Oracle Command Frames

In the experiments we describe in this paper, we perform a reduction on the basis of *oracle command frames*. The slots of an oracle command frame typically constitute a subset of the slots of the (usually over-specified) action frame and represents the semantic concepts that are actually expressed in the command, i.e. only frame slots that the participant refers to in the command, are filled in. Figure 3 shows the oracle command frame corresponding to the command "*put the jack of clubs on the queen of hearts*".

For some commands in the patience corpus, multiple mappings to frame slot values are possible. For instance, if a participant says "*put the black king in column three*", the word "*black*" refers to two possible values for the frame slot `from_suit`, i.e. `spades` or `clubs`. Therefore, this command has two oracle command frames: one version with `from_suit=spades` and one version with `from_suit=clubs`. In such cases, multiple corresponding oracle command frames are added.

2.3. Transcription and Annotation

In the next phase, orthographic transcriptions of the audio commands were created manually. In addition, the transcriptions were manually annotated using a concept tagging approach. This means that each command is segmented into chunks of words, which are tagged with the semantic concepts to which they refer. The concepts are, in this case,

Tag	Corresponding Frame (Slot)
I_FS	movecard(from_suit)
I_FV	movecard(from_value)
I_FF	movecard(from_foundation)
I_FC	movecard(from_fieldcol)
I_FH	movecard(from_hand)
I_TS	movecard(target_suit)
I_TV	movecard(target_value)
I_TF	movecard(target_foundation)
I_TFE	movecard(target_foundationempty)
I_TC	movecard(target_fieldcol)
I_TCE	movecard(target_fieldcolempty)
I_DC	dealcard()
O	-

Table 1: The set of concept tags used for annotation.

Leg	**de**	**klaveren**	**boer**	**op**	**de**	**harten**	**vrouw**
Put	the	clubs	jack	on	the	hearts	queen
O	O	I_FS	I_FV	O	O	I_TS	I_TV

Figure 4: Example of a command transcription annotated with concept tags.

slots in the frame-based description of the associated action, or, if the associated action frame does not contain any slots, the complete action frame. Thus, in the context of the patience game, the set of concept tags consists of the slots in the movecard action frame, plus one concept tag for the dealcard frame.

We use a tagging framework which is based on so-called IOB tagging, commonly used in the context of phrase chunking tasks (Ramshaw and Marcus, 1995). Words inside a chunk are labeled with a tag starting with I and words outside the chunks are labeled with an O tag, which means that they do not refer to any concept in the action frame. In the traditional IOB tagging framework, words at the beginning of a chunk are labeled with a tag starting with B. However, these B tags are typically only useful to indicate chunk boundaries when multiple chunks of the same type are immediately adjacent to each other. This does not occur in our data, however, yielding the complete tag set, shown in Table 1. The annotation of the command of Figure 2 is illustrated in Figure 4.

A Look inside the Patience Corpus

In this subsection, we will highlight some typical features and idiosyncratic patterns that can be found in the patience corpus. Figure 5 shows the most frequently used movecard command structures. The most frequent movecard structure is a structure in which the suit and value of the from card and the to card are specified, as shown in Figure 5(a). There is a lot of lexical variation of the prepositions and the verbs which are used in this structure. In addition, the position of the verb may vary considerably. The most frequent positions are the first position (usually imperative (cf. Figure 5(a)) and the final position (in infinitive form (cf. Figure 5(b)), but it may also occur in the position following the I_FV element, as in "*de harten vijf mag op de klaveren zes*" (*the hearts five may [be put] on the clubs six*).

Most participants used a specific word or phrase to move a card to one of the foundation stacks, without specifying which stack. Two examples are shown in Figures 5(b) and 5(c). Some participants also used specific phrases to move a card to an empty foundation stack, such as in the commands shown in Figure 5(e→f). When moving a king to an empty column, most participants used the structure shown in Figure 5(d). One participant, however, used the word *"afleggen"* ("*lay down*") for this purpose, which other participants typically used to express <to_foundation>, as shown in Figure 5(c). This type of inter-speaking variation underlines the importance of the adaptability of the ALADIN approach: a flexible C&C interface should adapt to the idiosyncracies of specific users and should not predefine the vocabulary and grammar with which the device is to be manipulated.

The column numbers and foundation stack numbers were rarely specified in the commands. Some participants referred to column numbers when moving a king to an empty column. Figure 6(a) shows an example. There were also some participants, however, who referred to specific ranges of columns or foundation stacks, by using the words "*links*" (*left*) and "*rechts*" (*right*). An example is shown in Figure 6(b). In this case, the word "*links*" ambiguously refers to column numbers 1, 2 and 3.

Figure 6(b) also shows another phenomenon, which occurred frequently: the use of the word "*zwarte*" (*black*), referring to the suits clubs and spades (and, similarly, the word "*rode*" (*red*), referring to the suits hearts and diamonds). Both in the black/red situation and the left/right situation, one single word refers to a range of possible frame slot values. This means that the command expresses multiple options with respect to certain slot values: in case of the command in Figure 6(b), regarding the values of the slots <from_suit>, <from_column> and <to_suit>. As previously mentioned, this means that the command yields multiple oracle command frames (Section 2.2), in which all possible combinations of values within the specified ranges are represented.

Another interesting phenomenon occurred in some cases, when a pile of multiple cards was moved from one column to another. In such cases, some participants specified all cards to be moved - an example is shown in Figure 6(c) - or the first and the last card in the pile to be moved. As shown in Figure 6(c), only the highest card in the pile is labeled with the concept tags I_FS and I_FV; the other cards are not represented in the frame description (and do not need to be).

Especially during the first few games, many participants showed some development with respect to the command structures that were used. Participants tended to shorten their commands as the games progressed, by, for instance, leaving out determiners and verbs, and sometimes even the card suits. In addition, the command structures of some of the participants gradually became more stable over the course of the games. It seems that many participants needed some time to establish the command structures that worked best for them. This type of intra-speaker variation over time

(a)	**[leg*]**	**[de]**	**harten**	**vijf**	**op**	**[de]**	**klaveren**	**zes**
	[put*]	[the]	hearts	five	on	[the]	clubs	six
	`[O*]`	`[O]`	`I_FS`	`I_FV`	`O`	`[O]`	`I_TS`	`I_TV`

(b)	**[de]**	**schoppen**	**drie**	**naar**	**boven**	**[plaatsen*]**
	[the]	spades	three	to	top	[move*]
	`[O]`	`I_FS`	`I_FV`	`O`	`I_TF`	`[O*]`

(c)	**[de]**	**klaveren**	**twee**	**afleggen**	**[bovenaan]**
	[the]	clubs	two	lay-down	[at-the-top]
	`[O]`	`I_FS`	`I_FV`	`I_TF`	`[I_TF]`

(d)	**[de]**	**harten**	**koning**	**naar**	**[de]**	**lege**	**plaats**
	[the]	hearts	king	to	[the]	empty	space
	`[O]`	`I_FS`	`I_FV`	`O`	`[O]`	`I_TCE`	`I_TCE`

(e)	**[de]**	**ruiten**	**aas**	**naar**	**het**	**groene**	**vak**
	[the]	diamonds	ace	to	the	green	field
	`[O]`	`I_FS`	`I_FV`	`O`	`O`	`I_TFE`	`I_TFE`

(f)	**[de]**	**klaveren**	**aas**	**op**	**een**	**leeg**	**vakje**	**bovenaan**	**[leggen*]**
	[the]	clubs	ace	on	an	empty	field	at-the-top	[put*]
	`[O]`	`I_FS`	`I_FV`	`O`	`O`	`I_TFE`	`I_TFE`	`I_TFE`	`[O*]`

Figure 5: The most frequently used `movecard` command structures, ranked according to frequency of occurrence. Optional words and tags are shown in [ ]. * indicates that the position of the verb varies.

(a)	**de**	**schoppen**	**koning**	**op**	**de**	**tweede**	**plaats**
	the	spades	king	on	the	second	position
	`O`	`I_FS`	`I_FV`	`O`	`O`	`I_TC`	`I_TC`

(b)	**de**	**zwarte**	**vier**	**links**	**naar**	**de**	**zwarte**	**drie**
	the	black	four	on-the-left	to	the	black	three
	`O`	`I_FS`	`I_FV`	`I_FC`	`O`	`O`	`I_TS`	`I_TV`

(c)	**leg**	**harten**	**vier**	**en**	**schoppen**	**drie**	**op**	**klaveren**	**vijf**
	put	hearts	four	and	spades	three	on	clubs	five
	`O`	`I_FS`	`I_FV`	`O`	`O`	`O`	`O`	`I_TS`	`I_TV`

Figure 6: Examples of more unusual command structures.

is again an important point of reference in the context of the ALADIN approach: the system should adapt over time to changes in the user's linguistic behavior.

3. Concept Tagging: Proof-of-the-Principle Experiments

The sequence tagging approach, illustrated in Figures 4, 5 and 6, presents a decidedly different type of representation, compared to traditional context-free grammar approaches (Figure 1), since no grammar in the traditional sense of the word is being produced. In that respect, our approach also differs from recent approaches in which context-free grammars constitute at least a part of the grammar framework, such as described in Starkie (2001) and in Wang and Acero (2005; 2006). The idea behind the sequence tagging approach is in fact more akin to that coined in Hahn et al. (2008), although this research effort does not directly refer to grammar induction as such.

In terms of grammars and parsing, we might dub our concept-tagging approach *shallow grammar induction*: similar to the technique of *shallow parsing* (Ramshaw and Marcus, 1995; Daelemans et al., 1999), we speculate that we do not need to construct a complete parse tree to enable successful processing of the data, but rather that a shallow representation of the syntactic/semantic compositionality of the utterance can suffice.

Furthermore, the final grammar module in the ALADIN system will need to be able to automatically induce these concept tags. Whereas context-free grammars have been proven to be very hard to automatically induce (de Marcken, 1999; Klein, 2005), particularly on the basis of limited training data (De Pauw, 2005), encouraging results have been reported in the unsupervised induction of sequence tags (Collobert et al., 2011). Furthermore, in contrast to traditional unsupervised grammar induction approaches that only work on the basis of raw data, we have additional pseudo-semantic information at our disposal in the form of action frames, that further help streamline the weakly supervised induction process.

In this section, we will describe the experimental setup for *supervised* concept tagging of the patience C&C task. These experiments serve as a proof-of-the-principle experiment that showcases the *learnability* of the task in optimal conditions, particularly in terms of the minimally required amount of training data needed to bootstrap successful concept tagging. In these experiments, the annotated corpus is used as training material for a data-driven tagger, which is subsequently used to tag previously unseen data. As our

	Baseline		Optimized	
	Mean	SD	Mean	SD
Tag accuracy (%)	77.8	4.0	96.9	1.3
Chunk accuracy ($F_{\beta=1}$)	73.8	3.4	96.5	1.3

Table 2: Ten-Fold Cross Validation: Experimental Results

tagger of choice, we opted for MBT, the memory-based tagger (Daelemans and van den Bosch, 2005; Daelemans et al., 2010).

3.1. Ten-fold Cross Validation

We tested the overall generalization capability of the tagger on the patience data, by performing a ten-fold cross-validation experiment on the complete data set of 2020 utterances. Each utterance in the data set was randomly assigned to one of ten sub-samples. Ten experiments were performed, each time using a different sub-sample as the evaluation set, with the remaining nine folds as the training set, including one development set to perform feature optimization. This means that the system is being trained and evaluated on utterances from different users.

The metrics used for the evaluation of the concept tagging performance are the tag accuracy and the chunk accuracy. The tag accuracy is the ratio of the number of correctly predicted tags; the chunk accuracy is the F-score for correctly predicted chunks, which means that the concept tags, as well as the borders of the predicted chunks are included in the evaluation. The accuracies with an optimized set of features[5] were compared to the accuracies in a baseline condition, in which only the focus word was used as a feature (and thus no context information was used).

The mean tag and chunk accuracies in the ten-fold cross-validation experiments are shown in Table 2. The mean tag accuracy with the optimized feature set is 96.9%, and the mean chunk accuracy is a bit lower at 96.5%. In the baseline condition, the mean tag accuracy is 77.8% and the mean chunk accuracy is 73.8%. The relatively large gap between the tag and chunk accuracies in the baseline condition is probably caused by the lack of coherence in that condition. Since no context features were used for tagging, chunk accuracies are lower.

3.2. Learning Curves

In the targeted ALADIN application, the number of utterances used to train the system, should be as small as possible, i.e. the training phase should be as brief as possible in order to limit the amount of extraneous physical work or assistance needed for training by the physically impaired person. In order to get an idea of the minimal number of training utterances needed to enable successful concept tagging, we evaluated the supervised tagging performance with increasing amounts of training data, resulting in learning curves.

In the learning curve experiments, we tried to mimic the ALADIN learning situation as much as possible. For each participant, a separate learning curve was made, since the learning process in the targeted ALADIN application will be personalized as well. For each learning curve, the last fifty utterances of a participant were used as a constant test set. The remaining utterances of the same participant were used as training material. The chronological order of the commands, as they were uttered by the participant, was preserved, in order to account for processes regarding the development of the users' command structure and vocabulary use during the games. In each experiment, the first k utterances were used as training data, k being an increasing number of slices of ten utterances. The feature set used by the tagger, was optimized in advance by means of a development set, consisting of the last 25% of the training data.

Figure 7 displays the learning curves for tag accuracies and chunk accuracies. There is a lot of variation between the participants in accuracy using the first 100 training utterances. For all participants, except participant 6, the tag accuracy reaches 95% or more with 130 training utterances, and levels off after that. The chunk accuracies tend to be slightly lower, but six out of eight curves still reach at least 95% chunk accuracy around 130 utterances. For two participants, the accuracies go up to 100%, with training set sizes of respectively 40 and 100 utterances. The baseline accuracies, averaged over all participants, are also shown in Figure 7. The maximum tag and chunk accuracies reached on average in the baseline condition, are 79.8% and 75.6% , respectively (using 170 training utterances). Most individual learning curves with optimized features are far above the average baseline curve, except the trailing curve of participant 1.

The sudden leap in this curve between 80 and 90 training utterances is due to the introduction of a new utterance for the `dealcard` move by the participant after the 80th utterance. After that, the participant kept using this new utterance (consisting of two previously unseen words), and in the test data, that same utterance occurred frequently as well. Until it was encountered in the training data, this utterance could not be successfully tagged with the appropriate concepts.

The fact that the tag accuracy for participant 6 remains relatively low (maximum around 92%), is mainly due to a rather high level of inconsistency and ambiguity in the command structures that were used. One remarkable source of errors in this case is a structure repeatedly occurring in the test set and occurring only twice in the largest training set. This is particularly difficult to learn: a structure in which multiple cards to be moved (in one pile) are specified, such as in "*de rode twee, de zwarte drie, de rode vier en de zwarte vijf naar de rode zes*" (*the red two, the black three, the red four and the black five to the red six*). In such cases, only the highest card of the moved pile (*black five* in the example) should be labeled with `I_FS` and `I_FV` tags (since only that card is represented in the action frame) and the lower cards (*red two*, *black three* and *red four*) should be tagged with `O` tags. Many errors were made in the tag-

[5]Feature selection was performed on the basis of a development set. MBT can use disambiguated tags (left context), words (left/right context) and ambiguous tags (for the focus word and right context) as features. Morphological features to disambiguate unknown words were not considered, since these will not be available to the final ALADIN system either.

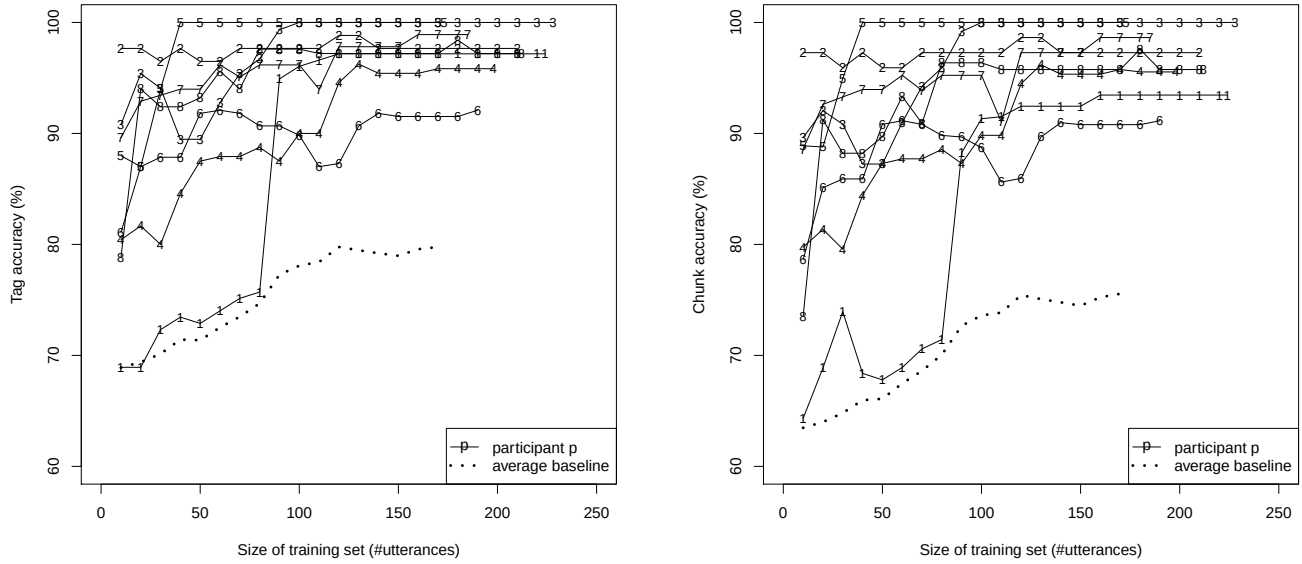


Figure 7: Learning curves viz. tag accuracy (left) and chunk accuracy (right). The solid curves show the accuracies per participant in the condition with the optimized feature set. The dotted curve shows the accuracies in the baseline condition, averaged over all participants.

ging of this type of structure. An example of ambiguity in the commands is the use of the phrase "*groene vak*" (*green field*) for both an empty foundation stack and an empty column.

The commands of participants 2 and 5 were structurally very consistent throughout the games, resulting in very fast learning. The learning curve of participant 5 reaches a tag accuracy of 100% using as little as forty training utterances. The curve of participant 2 immediately starts with an extremely high accuracy of 97.7% using only ten training utterances. However, it does not reach 100%, mainly due to the presence of a restart confusing the tagger "*schoppen boer op schoppen... op... schoppen boer op harten vrouw*" (*clubs jack on clubs... on... clubs jack on hearts queen*) and one clear inconsistency: using the phrase "*naar boven*" (*up*) to move a king to an empty column, whereas this phrase was previously only used for moving a card to the foundation.

The curve of participant 3 does reach 100% accuracy, but has a remarkable dip in the beginning of the curve. This is due to the fact that in the utterance numbers 20 to 50, the specification of the suit was often dropped (e.g. "*de drie op de vier*" *the three on the four*), whereas in the utterances before and after that, the specification of the suit was often included, as well as in many of the test utterances.

3.3. Discussion

The learning curves in Figure 7 show that with around 130 training utterances, between 95% and 100% tag accuracy could be reached for all participants, except one. The chunk accuracies tend to be a bit lower, but six out of eight curves still reach between 95% and 100% chunk accuracy using around 130 training utterances. After 130 training utterances (in some cases even earlier) a plateau is usually reached, meaning that adding more utterances does not significantly improve the tagging performance any more. This implies that having a participant play around two games of patience and subsequently transcribing and annotating the utterances, would usually provide sufficient training material for training a memory-based concept tagger to tag new transcribed utterances of that same participant with reasonable accuracy. It seems that after about 100 to 130 utterances of training material, accurate execution of commands can indeed be expected.

The initial part of the learning curves, i.e. using small training sets, varies considerably among participants. In general, differences between participants regarding the individual learning curves can be attributed mainly to differences in the level of consistency and the level of ambiguity regarding the command structures and the words used. Disfluencies such as restarts have a negative effect on accuracy scores, especially those present in the test set.

The learning curves are all situated above the average baseline learning curve. This means, as expected, that in order to successfully attribute concept tags to words in patience commands, the use of the **context** of each word (not available to the unigram baseline tagger) is essential. Therefore, in the ALADIN application, a (shallow) grammar module is indeed needed in order to attribute a correct meaning to this type of commands.

The results of the ten-fold cross-validation experiments furthermore show that a memory-based concept tagger generalizes well over different sets of patience commands, with mean tag and chunk accuracies of 96.9% and 96.5%, respectively.

4. Conclusion & Future Work

This paper described a corpus of command & control utterances for the card game *patience*, as well as some preliminary experiments that gauge the feasibility of inducing this task automatically on the basis of training material. The patience corpus is a relevant case study for this type of research: while the language use is fairly constrained and the structural complexity is manageable, these utterances do require some kind of minimal detection of grammatical structure to trigger the intended controls.

The experimental results show that a *supervised* approach of concept tagging works very well for this task. The tenfold cross validation experiments show that state-of-the-art classification accuracy can be achieved on data spanning different users. The learning curve experiments performed for each user individually, mimic the intended training phase in the final ALADIN system. The experiments results show that after about 130 utterances have been processed by the system, a workable tag accuracy of 95% or more can be achieved. These results are encouraging and form a solid basis for further experimentation with unsupervised approaches and for the integration of the grammar module in a command & control domotica interface for people with a physical impairment.

In the final ALADIN system, the grammar induction module will work together with an acoustic word finding module that will identify which patterns in the acoustic signal correspond to which word candidates. These word candidates will need to trigger specific frame slots as well as their values (cf. bottom part of Figure 2). The grammar module will need to be able to deal with and help resolve ambiguities and inaccuracies of the word finding module, as it will not have access to the unambiguous identity of the words in the utterance. In our next set of experiments, we will further approach the setup of the ALADIN system, by training on indices of lattices, output by the speech recognizer, rather than on the idealized situation of using orthographically transcribed utterances.

One of the biggest challenges we have yet to tackle in this research effort is to move from the supervised approach to an *unsupervised* approach, where we will need to match tags to words (or word candidates) without reference to annotated training material. To this end, we will look into unsupervised part-of-speech tagging approaches and investigate if and how they can adapted to this particular task.

Acknowledgments

This research was funded by IWT-SBO grant 100049 (ALADIN). We would like to thank the participants in the *patience* recording sessions for their kind help.

5. References

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12, pp. 2461–2505.

Daelemans, W. & van den Bosch, A. (2005). *Memory-Based Language Processing*. Cambridge, UK: Cambridge University Press.

Daelemans, W., Buchholz, S. & Veenstra, J. (1999). Memory-based shallow parsing. In M. Osborne & E. Tjong Kim Sang (Eds.), *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-99)*. Bergen, Norway: pp. 53–60.

Daelemans, W., Zavrel, J., van den Bosch, A. & Van der Sloot, K. (2010). MBT: Memory-based tagger, version 3.2, reference guide. Technical Report 10-04, University of Tilburg.

de Marcken, C. (1999). On the unsupervised induction of phrase-structure grammars. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann & D. Yarowsky (Eds.), *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Technology*, pp. 191–208: Kluwer Academic Publishers.

De Pauw, G. (2005). Agent-based unsupervised grammar induction. In M.P. Gleizes, G. Kaminka, A. Nowé, S. Ossowski, K. Tuyls & K. Verbeeck (Eds.), *Proceedings of the Third European Workshop on Multi-Agent Systems*. Brussels, Belgium, December, 2005: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, pp. 114–125.

Hahn, S., Lehnen, P., Raymond, C. & Ney, H. (2008). A comparison of various methods for concept tagging for spoken language understanding. In Nicoletta . Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Klein, D. (2005). *The Unsupervised Learning of Natural Language Structure*. Ph.D. thesis, Stanford University.

Noyes, J. & Frankish, C. (1992). Speech recognition technology for individuals with disabilities. *Augmentative and Alternative Communication*, 8(4), pp. 297–303.

Ramshaw, L.A. & Marcus, M.P. (1995). Text chunking using transformation-based learning. In *Proceedings of the Third ACL Workshop on Very Large Corpora*.

Starkie, B. (2001). Programming spoken dialogs using grammatical inference. In *Advances in Artificial Intelligence, 14th Australian Joint Conference on Artificial Intelligence*. Springer Verlag.

Wang, Y. & Acero, A. (2005). SGStudio: Rapid semantic grammar development for spoken language understanding. In *Proceedings of Ninth European Conference on Speech Communication and Technology*. Lisbon, Portugal: International Speech Communication Association.

Wang, Y. & Acero, A. (2006). Rapid development of spoken language understanding grammars. *Speech Communication*, 48(3-4), pp. 390–416.

Bermuda, a data-driven tool for phonetic transcription of words

Tiberiu Boroş, Dan Ştefănescu, Radu Ion
Research Institute for Artificial Intelligence, Romanian Academy (RACAI)
Calea 13 Septembrie, nr. 13, Bucureşti, România
E-mail: {tibi, danstef, radu}@racai.ro

Abstract

The article presents the Bermuda component of the NLPUF text-to-speech toolbox. Bermuda performs phonetic transcription for out-of-vocabulary words using a Maximum Entropy classifier and a custom designed algorithm named DLOPS. It offers direct transcription by using either one of the two available algorithms, or it can chain either algorithm to a second layer Maximum Entropy classifier designed to correct the first-layer transcription errors. Bermuda can be used outside of the NLPUF package by itself or to improve performance of other modular text-to-speech packages. The training steps are presented, the process of transcription is exemplified and an initial evaluation is performed. The article closes with usage examples of Bermuda.

Keywords: grapheme-to-phoneme, letter-to-sound, phonetic transcription, text-to-speech, data driven

1. Introduction

The last years have brought about a dramatic increase in the performance of human-computer interaction tools and techniques. This has naturally led to their successful application in Information-Technology and related fields. Consequently, accessibility to digital resources for elderly or disabled people is enabled by diverse methods such as better text organization and navigation, improved text input methods or better text reading using text-to-speech tools.

We present the *Natural Language Processing Unified Framework* (NLPUF) for text-to-speech (TTS) synthesis, which is part of the deliverables within the METANET4U project[1]. It comprises of a set of NLP tools and a speech synthesis module that can all be used together or as standalone packages. Its functionality consists of text normalization, phonetic transcription, homograph disambiguation, prosodic synthesis and speech synthesis, each of the functions being performed by different tools in the package. The speech synthesis component uses concatenative unit selection and can be easily integrated with other speech synthesis engines such as MBROLA (Dutoit et al., 1996) or HTS (Zen et al., 2007).

NLPUF is under development at the moment, but it is nearing completion. Before a TTS system can synthetize voice starting from arbitrary text, certain tasks have to be performed by the Natural Language Processing (NLP) module of the TTS system. The NLP module deals with text normalization, phonetic transcription, prosody analysis etc. Text normalization refers to the expansion of acronyms, abbreviations, numeric or mathematical expressions, etc., while prosody analysis tries to learn how to mimic speech phenomena such as rhythm, stress and intonation starting from text (Huang et al., 2001).

In this paper we focus only on the *phonetic transcription* (PT) for out-of-vocabulary (OOV) words and the way PT can be used to improve text accessibility. The phonetic transcription of words can be obtained using lexicons for known or common words in a target language, but there will always be OOV words (technical terms, proper nouns etc.) regardless of the lexicon's size. In this situation, the system needs a method to predict OOV words' pronunciation. This is one of the fragile steps of the pre-processing and analysis of text, because errors produced by incorrect transcription predictions can accumulate with errors from other modules (this is known as *error propagation*) on the way to the speech synthesizer (the part of the TTS that is responsible for the actual voice synthesis), leading to misreads of the original message.

Also, presence of foreign words inside the text (a common issue in any type of text: news, novels, technical articles etc.) increases the complexity of the problem. Thus, phonetic transcription of OOV words would greatly benefit from language identification, which is still an unresolved problem for very short texts (da Silva and Lopes, 2006; Vatanen et al., 2010).

In the case of NLPUF, phonetic transcription of OOV words is performed by *Bermuda*, a data-driven tool that uses Machine Learning (ML) techniques to best fit a phonetic transcription given an input word. As any other ML approach, it uses features, which in this case are based solely on the letters and groups of letters within the input word. While using more context sensitive data (part of speech, syllabification etc.) may provide better results in some cases, we intend to show that state of the art results can be obtained without using such data. Such an application is therefore much faster and does not require additional resources. Moreover, homograph disambiguation is not an issue here. Bermuda deals only with OOV words, which means it is impossible to predict that such words have two or more pronunciations that distinguish between their senses. The task of homograph disambiguation can only be performed on known words and it is handled by a different component in our framework.

[1] www.metanet4u.eu

2. The role of phonetic transcription in improving text accessibility

Phonetic transcription (PT) has an important role in any TTS system. One of the objectives of speech synthesis from text is to allow the user to fully understand the message that is being transmitted. While prosody highly contributes to understanding the message, PT also has a notable impact. Incorrect PT can render an entire fragment meaningless and mispronunciation can lead to annoying results (e.g. the same word is mispronounced again and again in a phrase or paragraph) even if the information may sometimes be transmitted regardless of small erroneous transcriptions. PT errors can also add up to the prosody errors and have a negative impact on the overall system performance.

Spelling correction or query alteration also link to text accessibility when taking into account that most relevant information found on the Internet is written in languages of international use and not all users are native speakers of such languages. Research has shown the possibility of using phonetic similarity as a feature for spelling correction (Li et al., 2006). A misspelled word can be corrected by using its PT. Table 1 shows an example where a misspelled word and its correct form produce identical PTs.

	Word	Phonetic transcription
Correct	Conceive	k ax n s iy v
Incorrect	Conceiv	k ax n s iy v

Table 1: PTs for words “conceive” and “conceiv” produced by Bermuda

In section 8 we show another example where web query alteration can benefit from the PT of words.

3. Related Work

Phonetic transcription in terms of *letter-to-phoneme* conversion (L2P) can be a simple task for languages where the relationship between letters and their phonetic transcription is simple (languages that are preponderantly characterized by having phonemic orthography, e.g. Romanian) but for other languages it poses a set of challenges. For example, current state of the art systems for English phonetic transcription of OOV words have an accuracy of 65% to 71% when used on the CMUDICT dictionary (Jiampojamarn et al., 2008).

There are a series of different methods and approaches to L2P conversion from context sensitive grammars to using classifiers or techniques specific to part-of-speech tagging.

A notable example of using a context sensitive grammar for writing L2P rules (pertaining to English and French) is given by Divay and Vitale (1997), although nowadays automatically inducing L2P rules is the main route followed by mainstream L2P research.

The *Expectation-Maximization* (EM) algorithm (Dempster et al., 1977) (or variants of it) is used to find one-to-one or many-to-many alignments between letters and phonemes in (Black et al., 1998; Jiampojamarn et al., 2008; Paget et al. 1998). The main idea of this algorithm is that, certain pairs of letters and phonemes are much more frequent than others and EM is employed in an effort to automatically detect the most probable alignments given a list of pairs of words and their transcriptions as training data.

Another approach for PT uses Hidden Markov Models (HMMs). Given the L2P rules (i.e. the probability of a phoneme being generated by a letter and the probability of occurrence of a phoneme sequence), the problem of automatic PT can be restated as follows: find the optimum sequence of hidden states (phonemes) that account for the given observation (the OOV word that has been suitably segmented for this task). Research of this approach has been done by Taylor (2005) and Jiampojamarn et al. (2008). One interesting conclusion of their research is that more accurate results are achieved if the phonemic substrings are paired with letter substrings. The reason for this is that phonetic transcriptions are context dependent: at any given moment, the phoneme to be generated is dependent on the adjacent phonemes. Moreover, it also depends on a contextual window of letters of the given word (Demberg, 2007).

4. A general view on Bermuda

Bermuda implements 2 methods for the L2P conversion task. The first one employs a *Maximum Entropy* (ME) classifier (*PTC*) to predict the phonetic transcription of every letter in the context (word) and uses a set of features similar to the MIRA and Perceptron methods presented by Jiampojamarn et al. (2008). The second one uses *DLOPS* algorithm described in Boroş et al. (2012). Furthermore, each of the methods has been improved by employing another ME classifier (*ERC*) designed to correct common L2P errors made by these two methods. In addition to the features used by PTC, ERC uses new features based on the already predicted phonemes which have become available. In other words, Bermuda chains the first layer prediction (PTC or DLOPS) to a second layer ME classifier for error correction (ERC). This leads to an accuracy increase of 2% to 7%.

We aim to show how Bermuda can be used outside of the NLPUF package, as a stand-alone application, to improve performance in other modular TTS packages.

5. Phonetic Transcription as an Alignment Problem

All data-driven L2P systems require letter to phoneme alignment before a model for phonetic transcription can be created. This section presents a method for obtaining such an alignment that is easy to implement. PT can be viewed as a translation process from the written form of the word (the “source language”) to its phonetic representation (the “target language”) (Laurent et al. 2009). Because aligning between words and phonetic transcriptions is similar to training a translation model, it is possible to use a well-known tool, explicitly designed for this kind of task: GIZA++ (Och and Ney, 2003).

GIZA++ is a free toolkit designed for aligning items in a parallel corpus, often used in the Statistical Machine Translation (SMT) field. Given pairs of unaligned (at word level) source and target sentences, it outputs word alignments within each pair. GIZA++ treats the word alignment task as a statistical problem and, as such, it can be applied to other problems that can be described in similar terms. Rama et al. (2009) showed that GIZA++ can be successfully used to preprocess training data for letter to sound conversion systems.

6. Bermuda training

Before any phonetic transcription can be produced, the system has to be trained. Bermuda accepts two types of files (plain aligned files and GIZA++ output files) as input for the training process.

Each line in the plain aligned files contains a word paired with its PT. Every symbol or set of symbols used for either the encoding of the word (characters/letters) or the encoding of the PT (phonemes) are <SPACE> separated. The paired elements are separated by a <TAB> character. The number of tokens of the elements in each pair must be equal. The word characters, which in reality do not have a corresponding symbol in the PT, are marked with the empty phoneme: "-", designed to preserve the equality (lines 4 and 5 of figure 1). If one word character emits more than one corresponding symbol in the PT, the character "." is used to link together the symbols (line 5 of Figure 1). In some cases, in which more word characters participate in forming a single sound, it is standard practice to associate only the last letter of the word with the PT and assign the empty phoneme to the other letters.

```
a b a n d o n<TAB>ax b ae n d ax n
a b a s i c<TAB>ax b ey s ih k
a b a t e r<TAB>ax b ey t ax r
a b a t t e d<TAB>ax b ae - t ih d
a b u s e r<TAB>ax b y.uw z ax -
```

Figure 1: Plain text training file

One training method for Bermuda is by using the alignment output of the GIZA++ toolkit. We run GIZA++ for a primary letter to phoneme alignment with default parameters (10 iterations of IBM-1, HMM, IBM-3 and IBM-4 models). To do this, the data has to be split into two files, one corresponding to the words (source file) and the second one corresponding to their phonetic transcription (target file). Every word in the source file must be on a single line, and its letters have to be separated by <SPACE>. Every line in the source file has a corresponding line in the target file.

source.txt

```
f l u                    (line 1)
c a u s e                (line 2)
t w a s                  (line 3)
s h i r e                (line 4)
a b a n d o n            (line 5)
```

target.txt

```
f l uw                   (line 1)
k ao z                   (line 2)
t w oh z                 (line 3)
sh ia                    (line 4)
ax b ae n d ax n         (line 5)
```

Before running GIZA++ we make sure it is compiled to be used outside of the Moses MT Toolkit. The following two lines should run successfully on the source and target files:

```
plain2snt.out target.txt source.txt
GIZA++ -S uk_beep.src.vcb -T target.vcb
-C source_target.snt -p0 0.98 -o output
```

One frequent mistake that GIZA++ makes is the forced NULL alignment on the phonemic side. Since an unaligned phoneme must be generated by one of the close-by letters, we devised a simple correction algorithm that looks at the letters that emitted the previous and the next phonemes and links the unaligned phoneme to the letter with which it was most frequently aligned to. In case of ties, it chooses the letter on the left side. Let's take for example the word *absenteeism* (Figure 2). Between the phonetic symbols aligned to S and M that is 'Z' and 'M' respectively, we have the unaligned (or NULL aligned) symbol 'AH'. In this case, we correct the alignment by assigning the phoneme 'AH' to the letter "S" because 'AH' between 'Z' and 'M' was most frequently aligned with 'S' (next to 'M').

The correction algorithm also inserts the empty phoneme for every NULL aligned letter. In Figure 2, the letter at position 8 (bold font) does not emit any symbol and so, we insert the empty phoneme in the PT at the appropriate position.

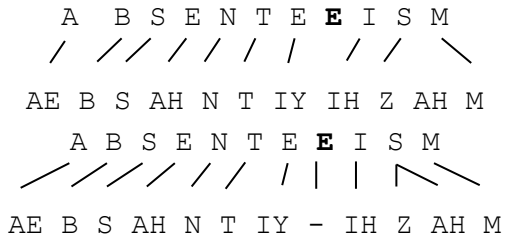


Figure 2: Alignment correction

Figure 3 represents an overview of our training process (comprising of letter to phoneme alignments and building models for the primary ME classifier, the DLOPS method and the second layer classifiers). DLOPS is a data-driven method used for generating PTs of OOV words by optimally adjoining maximal spans of PTs found in a given dictionary, corresponding to adjacent parts of the input word (Boroş et al., 2012). This is the case when GIZA++ is used for initial letter to phoneme alignment. If Bermuda receives plain text aligned files, the first two steps are ignored and Bermuda skips directly to training the first-layer methods. After the initial training of the first layer methods, Bermuda runs through the entire training corpora and produces PTs for every word using the two primary prediction methods (PTC and DLOPS). A new set of training data is compiled based on the

predictions made by the two methods and the real PTs in the training data. The second layer classifier (ERC) learns to correct the common mistakes of the two first-level methods, improving their accuracy.

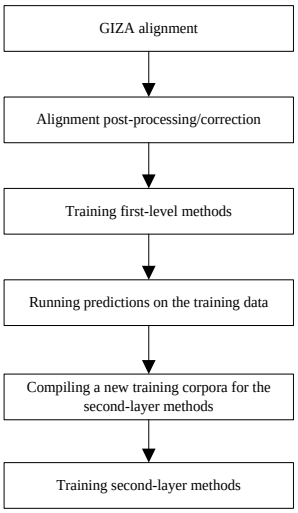


Figure 3*:* Training process

7. Usage and Testing

The current version of the system has been tested on two English dictionaries (BEEP UK – 250k words and CMUDICT US – 130k words) and on a Romanian dictionary extracted from the Romanian Speech Synthesis (RSS) Database (Stan et al., 2011). The training corpus was ten-folded and we ran Bermuda on every set while training on the other nine. The results show maximum performance for the PTC+ERC method as follows: CMUDICT 65%, the BEEP 71% and about 93% on the Romanian dictionary (the PT data for this dictionary has not been manually validated yet).

System	Word Accuracy on BEEP
PTC+ERC	71.31%
PTC	68.16%
DLOPS+ERC	66.40%
DLOPS	64.04%

Table 1: Word accuracy figures for the methods implemented by Bermuda

Table 1 shows an increase of about 2% to 3% in precision when chaining the second layer (ERC). These results are similar to those obtained by state of the art methods.
Once training files are available, Bermuda can be trained using the following lines:

```
bermuda.exe -gizatrain <giza A3 filename>
[-test]
bermuda.exe -plaintrain <plain aligned
filename> [-test]
```

If the *–test* option is specified, Bermuda splits the training corpora using the tenfold method. The data is divided into 10 files (folds), each having approximately 10% of the original corpus. After the split is performed, the tool shows the accuracy obtained on each of the 10 folds while sequentially training on the other 9. Accuracy is measured for each method in particular, so the user will be able to know which one to use in the final implementation.
The following command is used for running Bermuda:

```
bermuda.exe -run -m<1…4>
```

The second argument selects the method that will be used when predicting the PT of a given word. 1 corresponds to DLOPS method, 2 is used for PTC, 3 DLOPS+ERC and 4 means PTC+MRC. After the data for the specified method is loaded, the queries for the PT can be entered. Each letter must be space separated as in the following example:

```
Q:> a b s e n t e e i s m
       AE B S AH N T IY IH Z AH M 0.82%
       AE B S EH N T IY IH Z AH M 0.07%
       …
```

The example above displays results obtained using DLOPS method. This is the only method that currently shows the confidence level for each phonetic transcription variant.
Bermuda also has a custom evaluation method which takes as input a file with the same structure as the plain aligned training corpus and calculates its accuracy based on the data inside. This can be called using the following command:

```
bermuda.exe -customtest <filename>
```

8. Current state and future work

This tool is currently available for online testing and can be downloaded from RACAI's NLP Tools website[2]. The online version is trained for both Romanian (using a proprietary lexicon) and for English (using UK BEEP dictionary). It can produce phonetic transcriptions using any model specified (DLOPS, PTC, DLOPS+ERC or PTC+ERC). The phonetic representation is based on the symbols (e.g. "@" for the Romanian letter "ă") employed by each individual training lexicon, but we plan on mapping these symbols to the International Phonetic Alphabet (IPA) in order to have a unified phonetic transcription system. Referring back to section 2, IPA transcription could improve current query suggestion systems. For example, users would be able to enter queries based on their native perception of the pronunciation of words (write queries in their native language based on their phonetic perception). The system would then be able to map the PT to that of any other language, thus finding the correct spelling suggestion. We call this type of query input *perceptive search* and we plan on doing further research in this area as well. We need to mention that Bermuda can be used to map back phonemes to words by inversing the lexicon files, a task which implies a different technique in order to cope with homophones.

[2] http://nlptools.racai.ro/

9. Conclusions

We have presented a data-driven tool for L2P conversion, which is part of the NLPUF package but can also be used individually. Training and usage of this tool are fully covered in this paper.

Sections 2 and 8 show the role of phonetic transcription in improving text accessibility starting from its integration in TTS systems, spelling correction and/or alteration based on phonetic similarity and the possibility of using letter to phoneme conversion and phoneme to letter conversion for implementing perceptive search.

Our future plans include further development and fine-tuning work on the current methods and a complete set of tests for experimental validation using baselines provided by other L2P systems (e.g. using the same dictionaries as other systems). We also want to map the available dictionaries to IPA and to implement and test a perceptive search method based on Bermuda.

This tool will be free and available for download once the final tests are performed.

10. Acknowledgments

The work reported here was funded by the project METANET4U by the European Comission under the Grant Agreement No 270893.

11. References

Baayen, R., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database. In *Linguistic Data Consortium*, University of Pennsylvania, Philadelphia.

Black, A., Lenzo, K. and Pagel, V. (1998). Issues in building general letter to sound rules. In *ESCA Speech Synthesis Work-shop*, Jenolan Caves.

Boroş, T., Ştefănescu, D. and Ion, R. (2012). Data driven methods for phonetic transcription of words. In the *13th Annual Conference of the International Speech Communication Association* (submitted).

Bosch, A., and Canisius, S. (2006). Improved morpho phonological sequence processing with constraint satisfaction inference. In *Proceedings of the Eighth Meeting of the ACL-SIGPHON at HLT-NAACL*, pp. 41–49.

CMU. (2011). Carnegie Mellon Pronuncing Dictionary. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Content, A., Mousty, P., and Radeau, M. (1990). Une base de données lexicales informatisée pour le français écrit et parlé. In *L'Année Psychologique*, 90, pp. 551–566.

Da Silva, J. F., and Lopes, G. P. (2006). Identification of document language is not yet a completely solved problem. In *Proceedings of CIMCA'06*, pp. 212–219.

Dempster, A.P., Laird, N. M. and Rubin, D.B. (1977). Maximum likelihood from in-complete data via the em algorithm. In *Journal of the Royal Statistical Society*: Series B, 39(1), pp. 1–38.

Demberg, V. (2007). Phonological constraints and morphological preprocessing for grapheme-to-phoneme conversion. In *Proceedings of ACL-2007*.

Divay, M. and Vitale, A. J. (1997). Algorithms for grapheme-phoneme translation for English and French: Applications. In *Computational Linguistics*, 23(4), pp. 495–524.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F. and van der Vrecken, O. (1996). The MBROLA Project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes. In ICSLP'96, pp. 1393–1396.

Huang, X., Acero, A., and Hon, H. W. (2001). *Spoken Language Processing*. Upper Saddle River, NJ: Prentice-Hall.

Jiampojamarn, S., Cherry, C. and Kondrak, G. (2008). Joint processing and discriminative training for letter-to-phoneme conversion. In *Proceedings of ACL-2008: Human Language Technology Conference*, pp. 905–913, Columbus, Ohio.

Laurent, A., Deleglise, P. and Meignier, S. (2009). Grapheme to phoneme conversion using an SMT system. In Proceedings of the *10th Annual Conference of the International Speech Communication Association*.

Li, M., Zhang, Y., Zhu, M. and Zhou, M. (2006). Exploring distributional similarity based models for query spelling correction. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pp. 1025–1032.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, 29(1), pp. 19–51.

Pagel, V., Lenzo, K. and Black, A. (1998). Letter to sound rules for accented lexicon compression. In *International Conference on Spoken Language Processing*, Sydney, Australia.

Rama, T., Singh, A. K. and Kolachina, S. (2009). Modeling Letter-to-Phoneme Conversion as a Phrase Based Statistical Machine Translation Problem with Minimum Error Rate Training. In *Proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP 2009*, pp. 124–127, Suntec, Singapore.

Stan, A.,Yamagishi, J., King, S. and Aylett, M. (2011). The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate. In *Speech Communication*, 53 (3), pp. 442–450.

Taylor, P. (2005). Hidden Markov Models for grapheme to phoneme conversion. In *Proceedings of the 9th European Conference on Speech Communication and Technology*.

Vatanen, T., Jaakko Väyrynen, J. and Virpioja, S. (2010). Language Identification of Short Text Segments with N-gram Models. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation LREC'10*.

Zen, H., Nose, T., Yamagishi, J., Sako, S., and Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0. In *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, pp. 294–299.